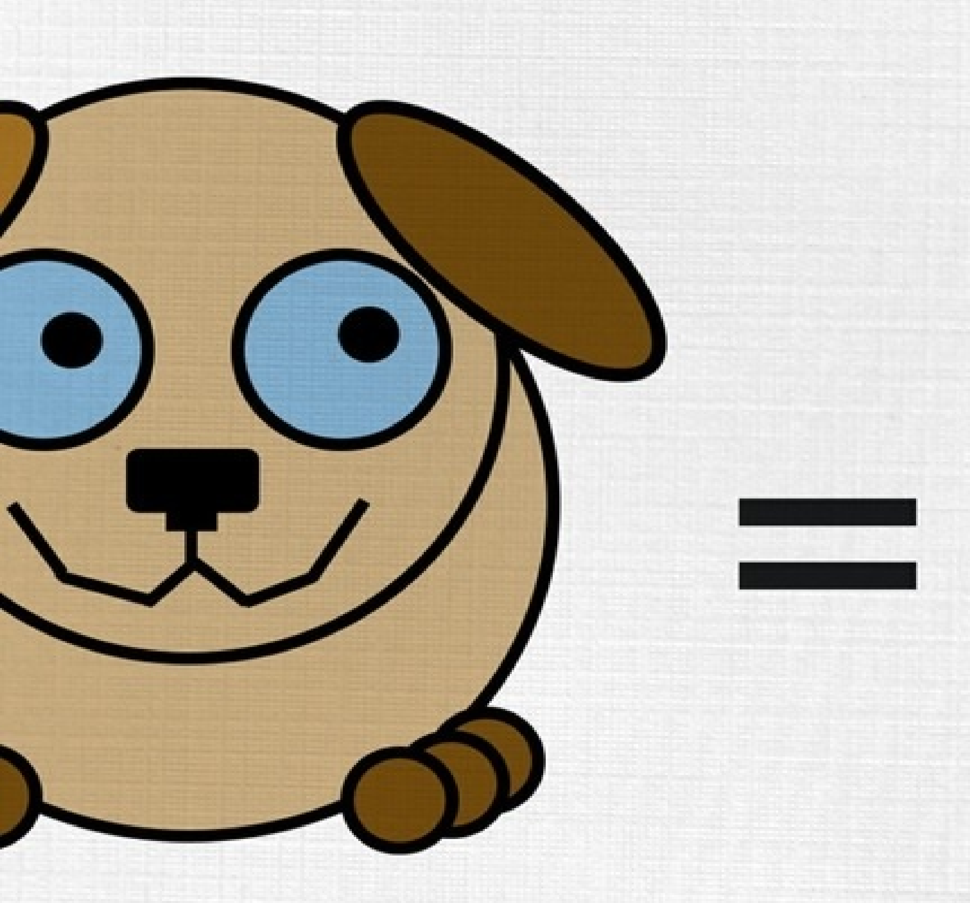
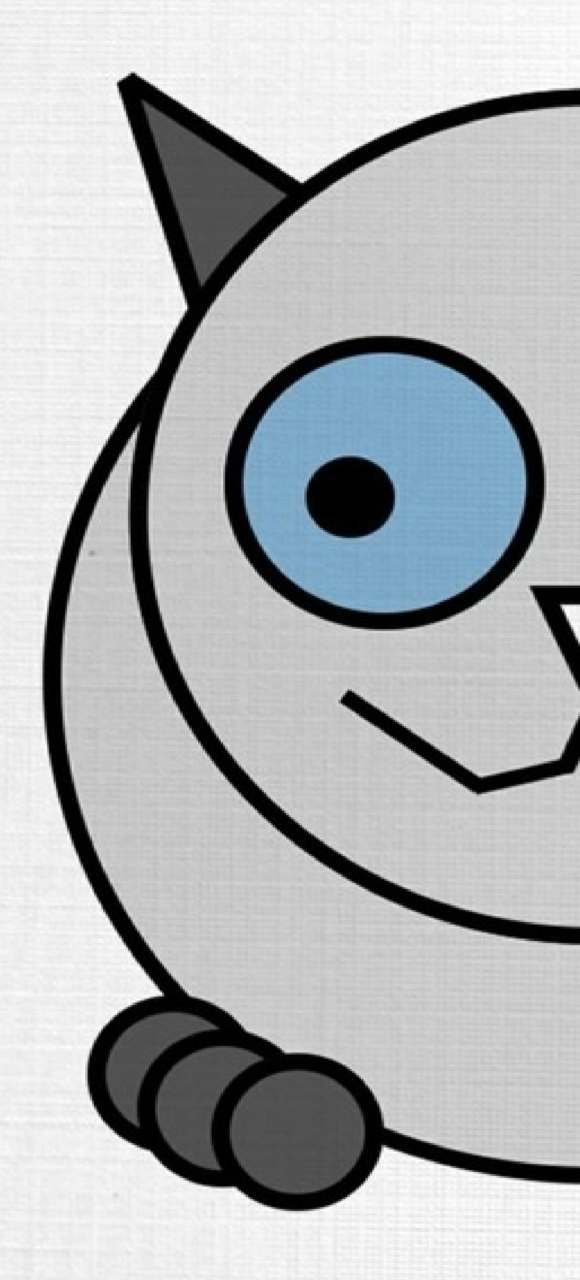
Владимир Савельев

Статистика
и котики

Самая дружелюбная книга
об анализе данных



=



При поддержке ЦИиР Юрия Корженевского

Annotation

Из этой книги вы узнаете, что такое дисперсия и стандартное отклонение, как найти t-критерий Стьюдента и U-критерий Манна-Уитни, для чего используются регрессионный и факторный анализы, а также многое и многое другое.

И все это — на простых и понятных примерах из жизни милых и пушистых котиков, которые дарят нам множество приятных эмоций.

Статистика и котики

Владимир Савельев

ISBN 978-5-4483-3995-0
Создано в интеллектуальной издательской системе Ridero

Предисловие

ОТ АВТОРА

Мало кто любит статистику.

Одни считают эту науку сухой и безжизненной. Другие боятся и избегают ее. Третьи полагают, что она бесполезна. Но у меня другое мнение на этот счет.

На мой взгляд, статистика обладает своей особой внутренней красотой. Ее можно увидеть, вглядываясь в корреляционную матрицу, рассматривая дендрограммы или интерпретируя результаты факторного анализа. За каждым статистическим коэффициентом стоит маленькое чудо, раскрывающее скрытые закономерности окружающего нас мира.

Но чтобы найти эту красоту, чтобы услышать поэзию, которая пронизывает статистику насквозь, необходимо преодолеть первоначальный страх и недоверие, вызванное внешней сложностью этого предмета.

Для того и написана эта книга. Чтобы показать, что статистика не такая страшная, как о ней думают. И что она вполне может быть такой же милой и пушистой, как котики, которые встретятся вам на страницах этой книги.

ОТ ПАРТНЕРА ИЗДАНИЯ

При слове «статистика» я вспоминаю британских ученых и выборы. Статистика — это многогранный инструмент. Иногда статистикой манипулируют, а можно открывать знания о реальном мире.

Автор написал книгу о базовой статистике в забавном формате. Старая система образования выдает порцию неинтересных и бесполезных знаний. А котики обучают, развлекая.

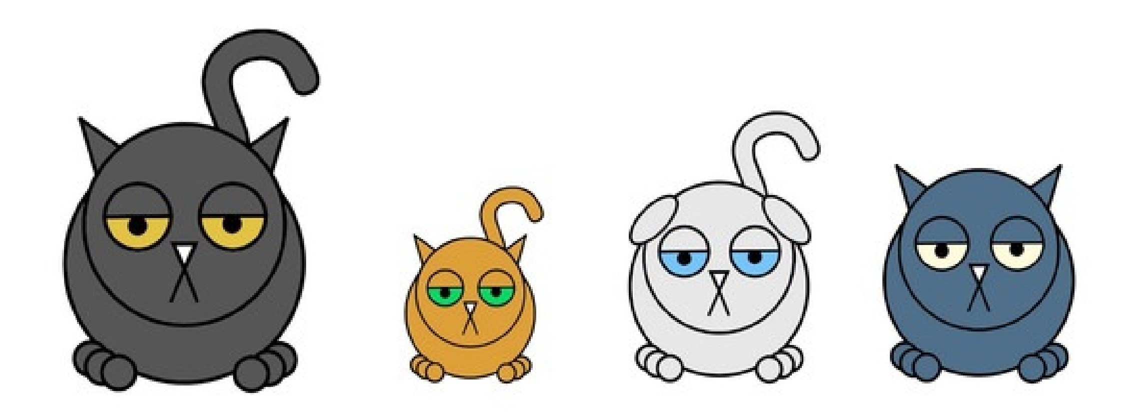
Когда мы изучаем данные, мы осознаем, что задача — найти соломинку в стоге иголок. И понять, сколько ещё стогов и соломы найдем дальше. Статистика в бизнесе помогает нам экономить деньги и открывать новые рынки. Экономия питает амбиции и потихоньку делает жизнь людей чуточку лучше.

Респект читателям. Респект автору.

Юрий Корженевский,
Центр Исследований и Разработки.
www.rnd.center

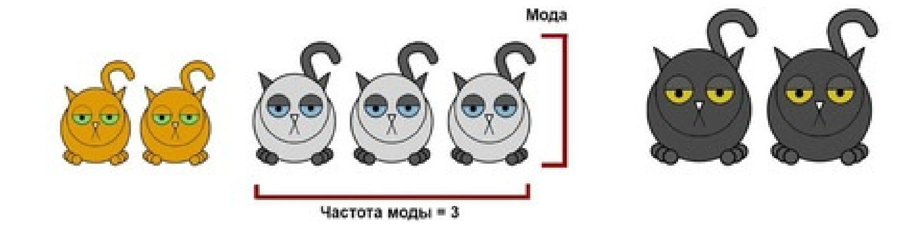
Глава 1.
Как выглядят котики
или основы описательной статистики

Котики бывают разные. Есть большие котики, а есть маленькие. Есть котики с длинными хвостами, а есть и вовсе без хвостов. Есть котики с висячими ушками, а есть котики с короткими лапками. Как же нам понять, как выглядит типичный котик?

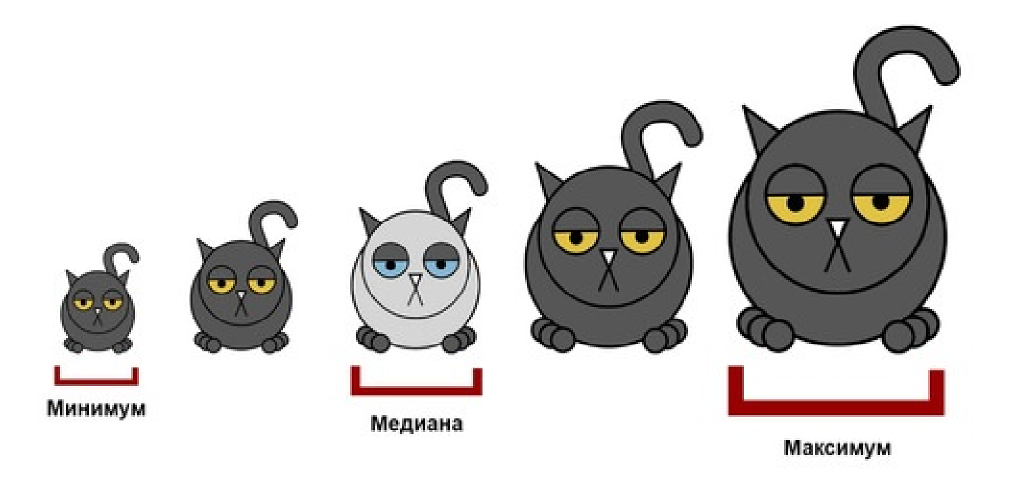


Для простоты мы возьмем такое котиковое свойство, как размер.

Первый и наиболее очевидный способ — посмотреть, какой размер котиков встречается чаще всего. Такой показатель называется *модой*.



Второй способ: мы можем упорядочить всех котиков от самого маленького до самого крупного, а затем посмотреть на середину этого ряда. Как правило, там находится котик, который обладает самым типичным размером. И этот размер называется *медианой*.



Если же посередине находятся сразу два котика (что бывает, когда их четное количество), то, чтобы найти медиану, нужно сложить их размеры и поделить это число пополам.

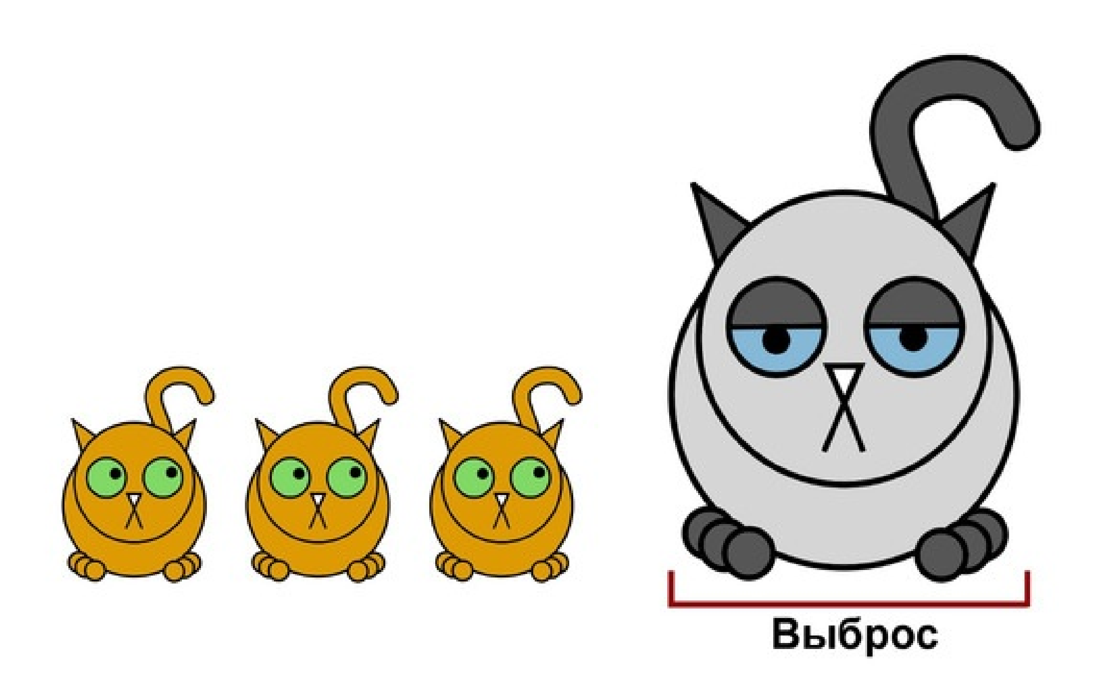


Последний способ нахождения наиболее типичного котика — это сложить размер всех котиков и поделить на их количество. Полученное число называется *средним значением*, и оно является очень популярным в современной статистике.



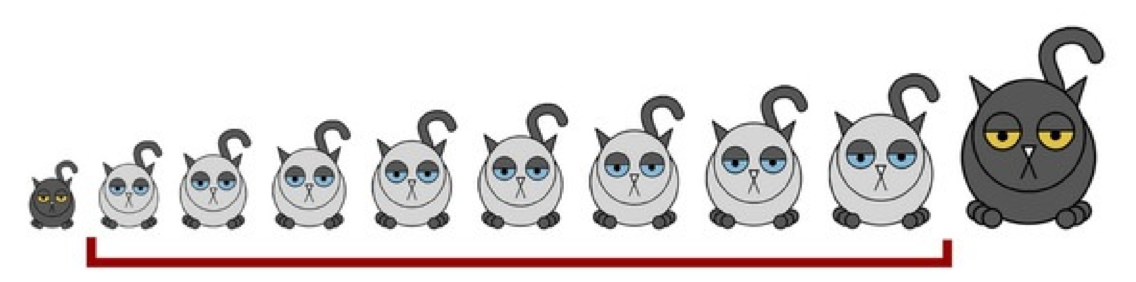
Однако, среднее арифметическое далеко не всегда является лучшим показателем типичности.

Предположим, что среди наших котиков есть один уникум размером со слона. Его присутствие может существенным образом сдвинуть среднее значение в большую сторону, и оно перестанет отражать типичный котиковый размер.



Такой «слоновый» котик, так же как и котик размером с муравья, называется *выбросом*, и он может существенно исказить наши представления о котиках. И, к большому сожалению, многие статистические критерии, содержащие в своих формулах средние значения, также становятся неадекватными в присутствии «слоновых» котиков.

Чтобы избавиться от таких выбросов, иногда применяют следующий метод: убирают по 5—10% самых больших и самых маленьких котиков и уже от оставшихся считают среднее. Получившийся показатель называют *усеченным (или урезанным) средним*.



Альтернативный вариант — применять вместо среднего медиану.

Итак, мы рассмотрели основные методы нахождения типичного размера котиков: моду, медиану и средние значения. Все вместе они называются *мерами центральной тенденции*. Но, кроме типичности, нас довольно часто интересует, насколько разнообразными могут быть котики по размеру. И в этом нам помогают меры изменчивости.

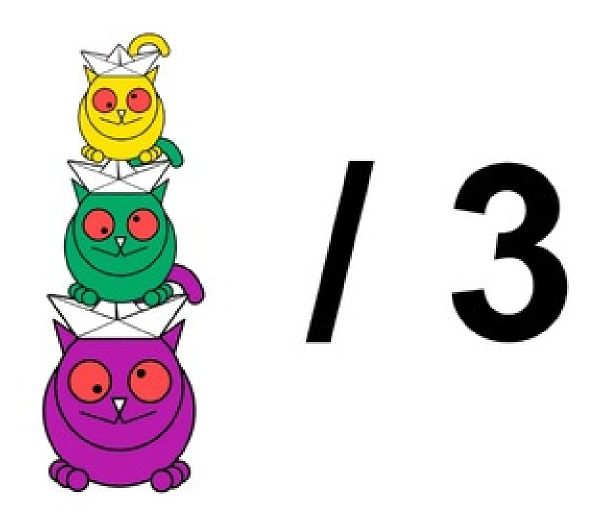
Первая из них — *размах* — является разностью между самым большим и самым маленьким котиком. Однако, как и среднее арифметическое, эта мера очень чувствительна к выбросам. И, чтобы избежать искажений, мы должны отсечь 25% самых больших и 25% самых маленьких котиков и найти размах для оставшихся. Эта мера называется *межквартильным размахом*.



Вторая и третья меры изменчивости называются *дисперсией* и *стандартным отклонением*. Чтобы разобраться в том, как они устроены, предположим, что мы решили сравнить размер некоторого конкретного котика (назовем его Барсиком) со средним котиковым размером. Разница (а точнее разность) этих размеров называется *отклонением*. И совершенно очевидно, что чем сильнее Барсик будет отличаться от среднего котика, тем больше будет это самое отклонение.



Логично было бы предположить, что чем больше у нас будет котиков с сильным отклонением, тем более разнообразными будут наши котики по размеру. И, чтобы понять, какое отклонение является для наших котиков наиболее типичным, мы можем просто найти среднее значение по этим отклонениям (т. е. сложить все отклонения и поделить их на количество котиков).



Однако если мы это сделаем, то получим 0. Для недоверчивых привожу доказательство:

$$\sum_{i=1}^{n} \Delta x_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n \sum_{j=1}^{n} \frac{x_j}{n} = \sum_{i=1}^{n} x_i - \sum_{j=1}^{n} x_j = 0$$

Это происходит, поскольку одни отклонения являются положительными (когда Барсик больше среднего), а другие — отрицательными (когда Барсик меньше среднего). Поэтому

необходимо избавиться от знака. Сделать это можно двумя способами: либо взять модуль от отклонений, либо возвести их в квадрат, который, как мы помним, всегда положителен. Последнее применяется чаще.



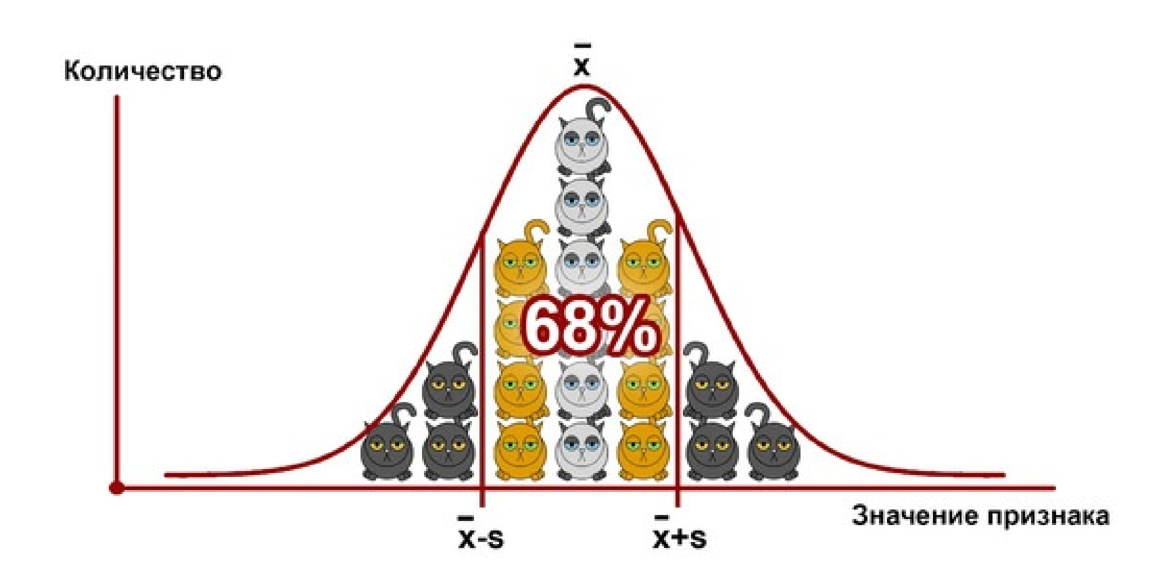
И, если мы найдем среднее от квадратов отклонений, мы получим то, что называется *дисперсией*. Однако, к большому сожалению, квадрат в этой формуле делает дисперсию очень неудобной для оценки разнообразия котиков: если мы измеряли размер в сантиметрах, то дисперсия имеет размерность в квадратных сантиметрах. Поэтому для удобства использования дисперсию берут под корень, получая по итогу показатель, называемый *среднеквадратическим отклонением*.



К несчастью, дисперсия и среднеквадратическое отклонение так же неустойчивы к выбросам, как и среднее арифметическое.

Среднее значение и среднеквадратическое отклонение очень часто совместно используются для описания той или иной группы котиков. Дело в том, что, как правило, большинство (а именно около 68%) котиков находится в пределе одного среднеквадратического отклонения

от среднего. Эти котики обладают так называемым *нормальным размером*. Оставшиеся 32% либо очень большие, либо очень маленькие. В целом же для большинства котиковых признаков картина выглядит вот так.

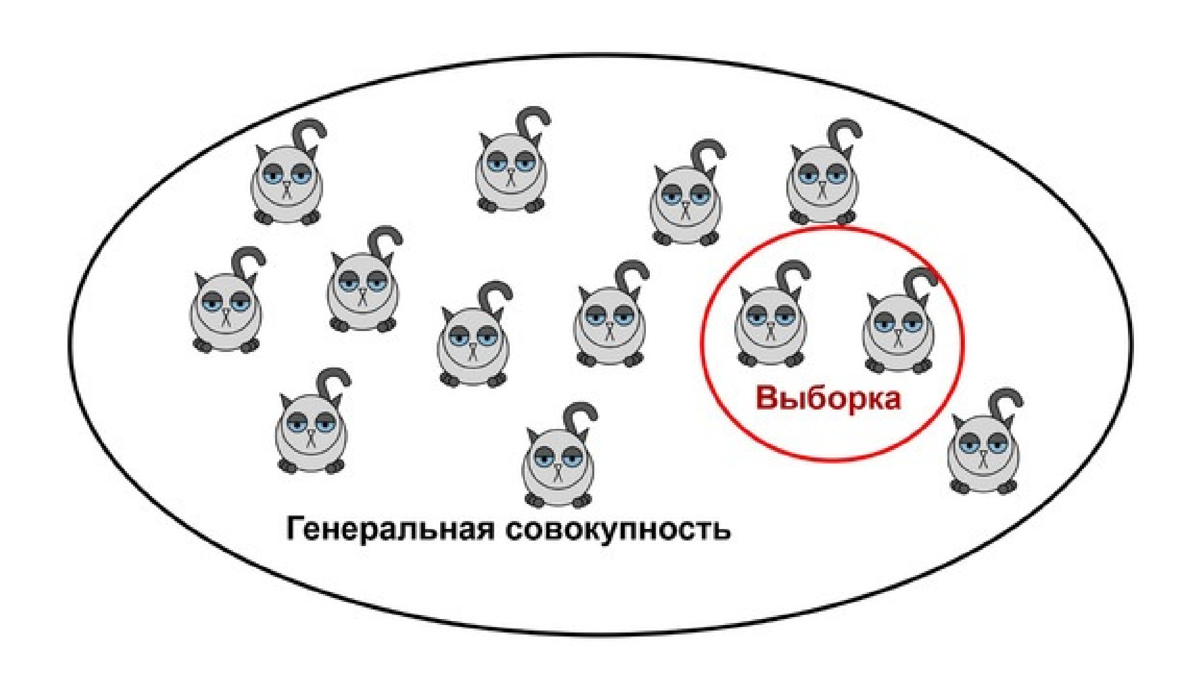


Такой график называется *нормальным распределением* признака.

Таким образом, зная всего два показателя, вы можете с достаточной долей уверенности сказать, как выглядит типичный котик, насколько разнообразными являются котики в целом и в каком диапазоне лежит норма по тому или иному признаку.

НЕМАЛОВАЖНО ЗНАТЬ!
Выборка, генеральная совокупность и два вида дисперсии

Чаще всего нас, как исследователей, интересуют все котики без исключения. Статистики называют этих котиков *генеральной совокупностью*. Однако на практике мы не можем замерить всю генеральную совокупность — как правило, мы работаем только с небольшим количеством котиков, называемым *выборкой*.



Очень важно, чтобы выборка была максимально похожа на генеральную совокупность.

Степень такой похожести называется *репрезентативностью*.

Необходимо запомнить, что существует две формулы дисперсии: одна для генеральной совокупности, другая — для выборки. В знаменателе первой всегда стоит точное количество котиков, а у второй — ровно на одного котика меньше.





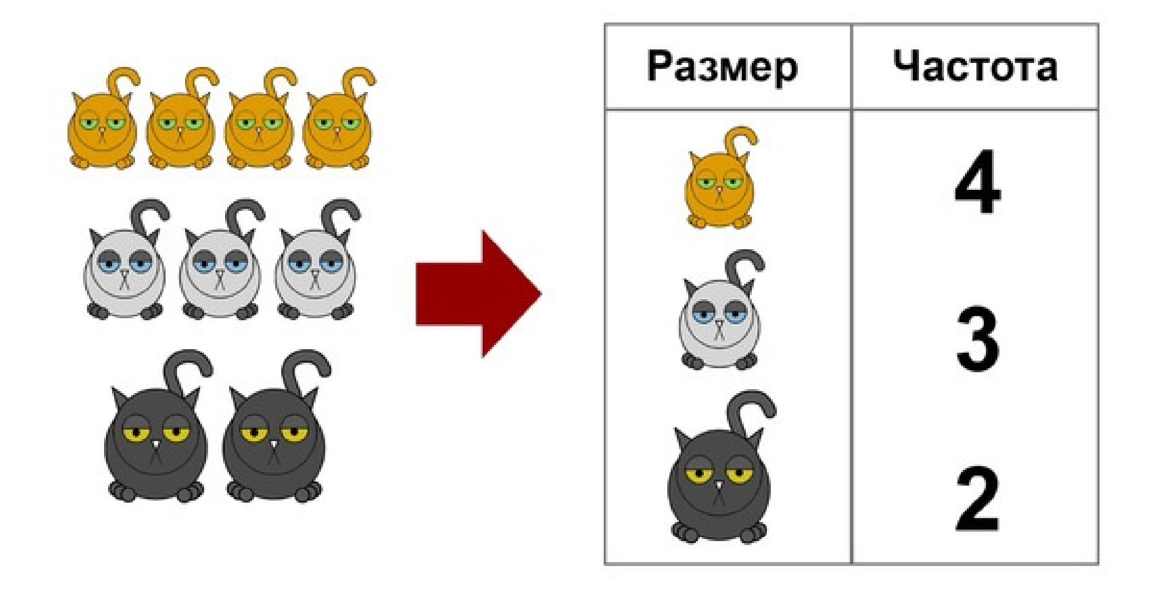
Корень из дисперсии генеральной совокупности, как уже было сказано, называется *среднеквадратическим отклонением*. А вот корень из дисперсии по выборке называется *стандартным отклонением*.

Однако не будет большой ошибкой, если вы будете пользоваться терминами *стандартное отклонение генеральной совокупности* и *стандартное отклонение выборки*. Чаще всего именно последнее и рассчитывается для реальных исследований.

Глава 2.
Картинки с котиками
или средства визуализации данных

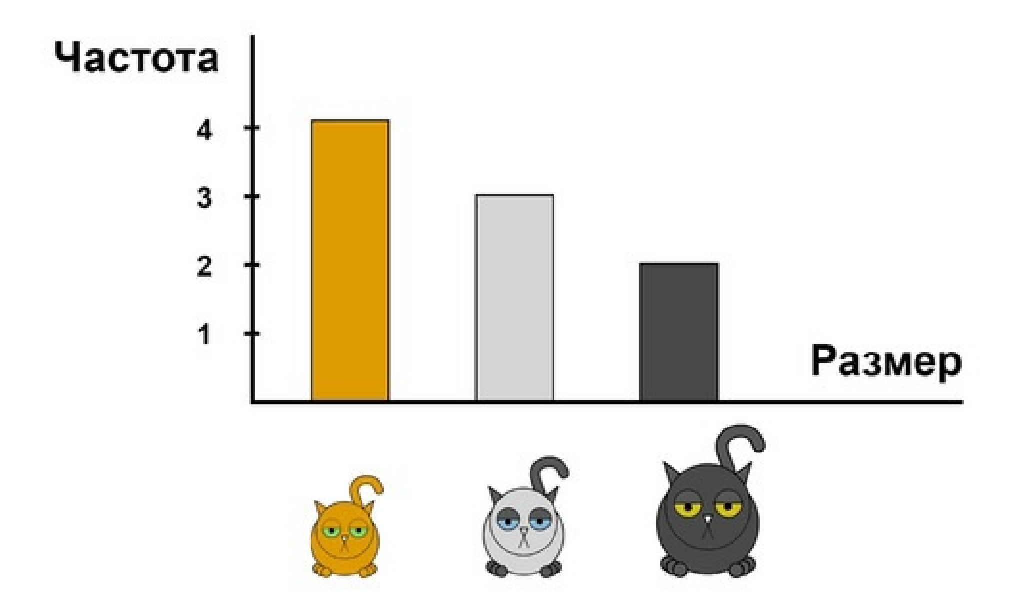
В предыдущей главе мы говорили про показатели, которые помогают определить, какой размер является для котиков типичным и насколько он бывает разнообразным. Но когда нам требуется получить более полные и зрительно осязаемые представления о котиках, мы можем прибегнуть к так называемым *средствам визуализации данных*.

Первая группа средств показывает, сколько котиков обладает тем или иным размером. Для их использования необходимо предварительно построить так называемые *таблицы частот*. В этих таблицах есть два столбика: в первом указывается размер (или любое другое котиковое свойство), а во втором — количество котиков при данном размере.

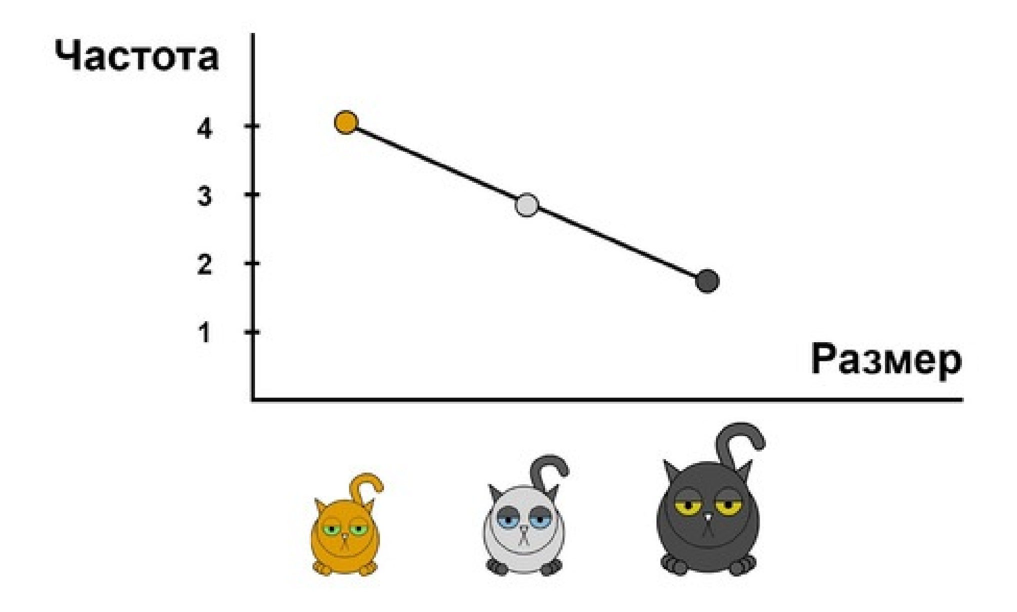


Это количество, кстати, и называется *частотой*. Эти частоты бывают *абсолютными* (в котиках) и *относительными* (в процентах).

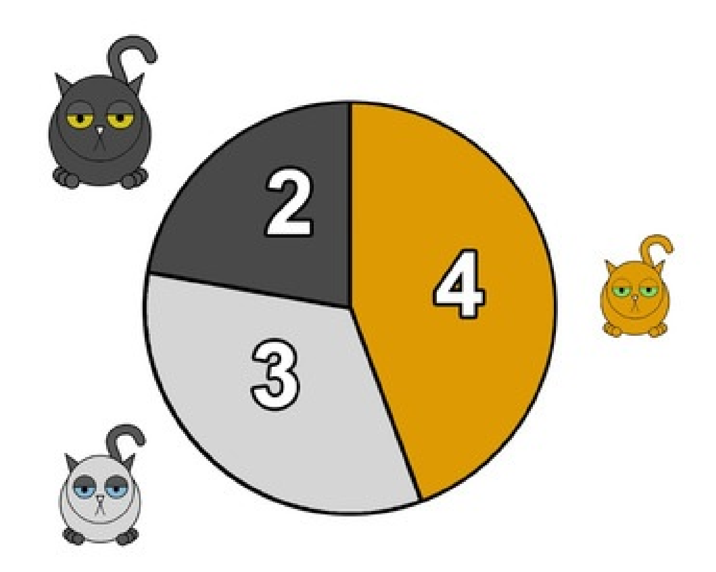
С таблицами частот можно делать много интересных вещей. Например, построить *столбиковую диаграмму*. Для этого мы откладываем две перпендикулярных линии: горизонтальная будет обозначать размер, а вертикальная — частоту. А затем — рисуем столбики, высота которых будет соответствовать количеству котиков того или иного размера.



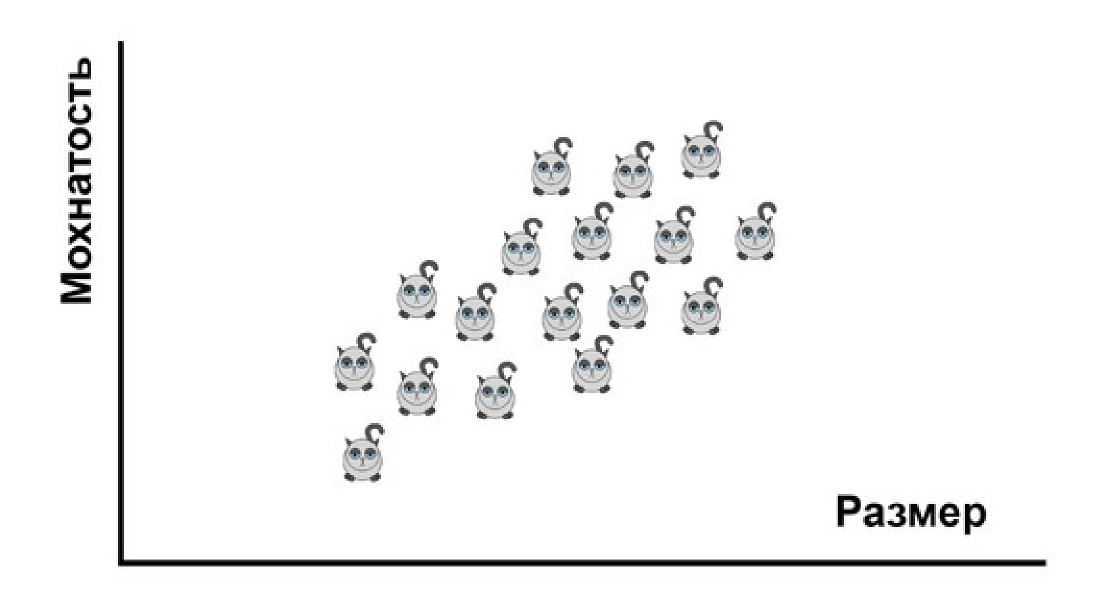
А еще мы можем вместо столбиков нарисовать точки и соединить их линиями. Результат называется *полигоном распределения*. Он довольно удобен, если котиковых размеров действительно много.



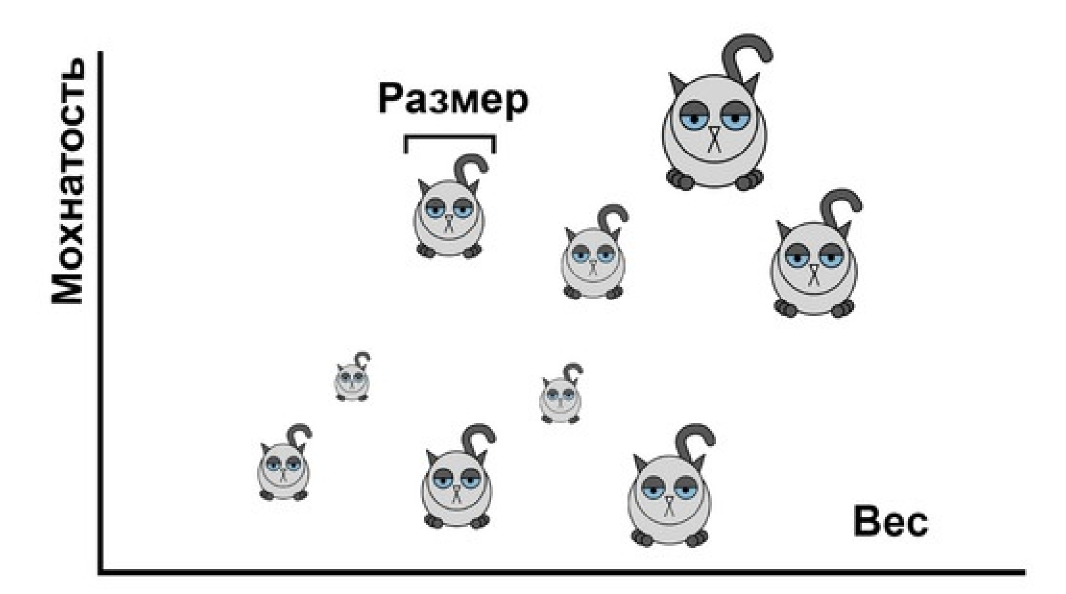
Наконец, мы можем построить *круговую диаграмму*. Величина каждого сектора такой диаграммы будет соответствовать проценту котиков определенного размера.



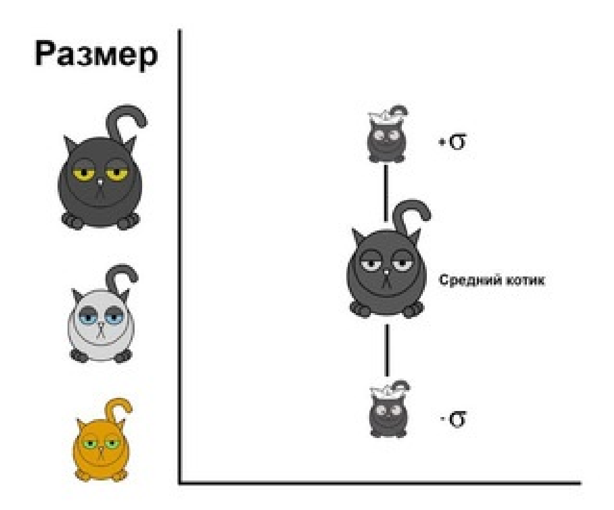
Следующая группа средств визуализации позволяет отобразить сразу два котиковых свойства. Например, размер и мохнатость. Как и в случае со столбиковыми диаграммами, первым шагом рисуются оси. Только теперь каждая из осей отображает отдельное свойство. А после этого каждый котик занимает на этом графике свое место в зависимости от степени выраженности этих свойств. Так, большие и мохнатые котики занимают место ближе к правому верхнему углу, а маленькие и лысые — в левом нижнем.



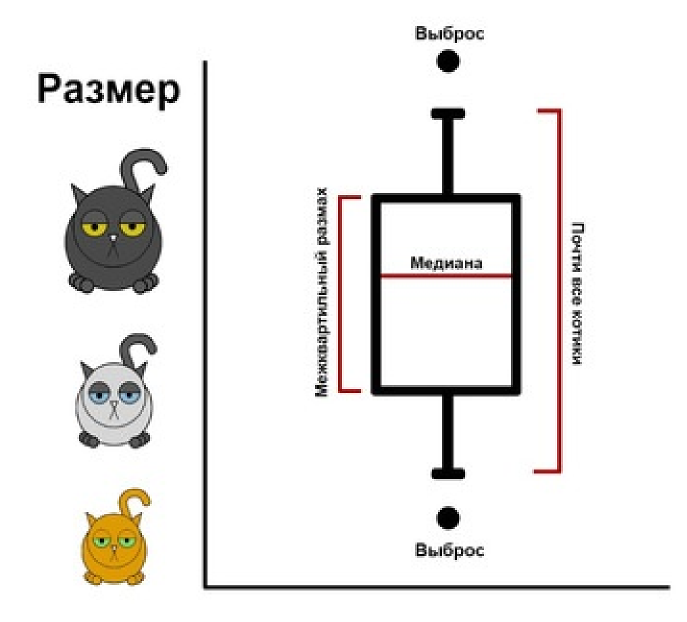
Поскольку обычно котики на данной диаграмме обозначаются точками, то она называется *точечной* (или *диаграммой рассеяния*). Более продвинутый вариант — *пузырьковая диаграмма* — позволяет отобразить сразу три котиковых свойства одновременно (размер, мохнатость и вес). Это достигается за счет того, что сами точки на ней имеют разную величину, которая и обозначает третье свойство.



Последняя крупная группа средств визуализации позволяет графически изобразить меры центральной тенденции и меры изменчивости. В простейшем виде это точка на графике, обозначающая, где находится средний котик, и линии, длина которых указывает на величину стандартного отклонения.



Более известным средством является так называемый *боксплот* (или *«ящик с усами»*). Он позволяет компактно отобразить медиану, общий и межквартильный размах, а также прикинуть, насколько распределение ваших данных близко к нормальному и есть ли у вас выбросы.



Помимо вышеперечисленных средств существует еще немало специфических, заточенных под определенные цели (например диаграммы, использующие географические карты). Однако, вне зависимости от того, какой тип диаграмм вы хотели бы использовать, существует ряд рекомендаций, которые желательно соблюдать.

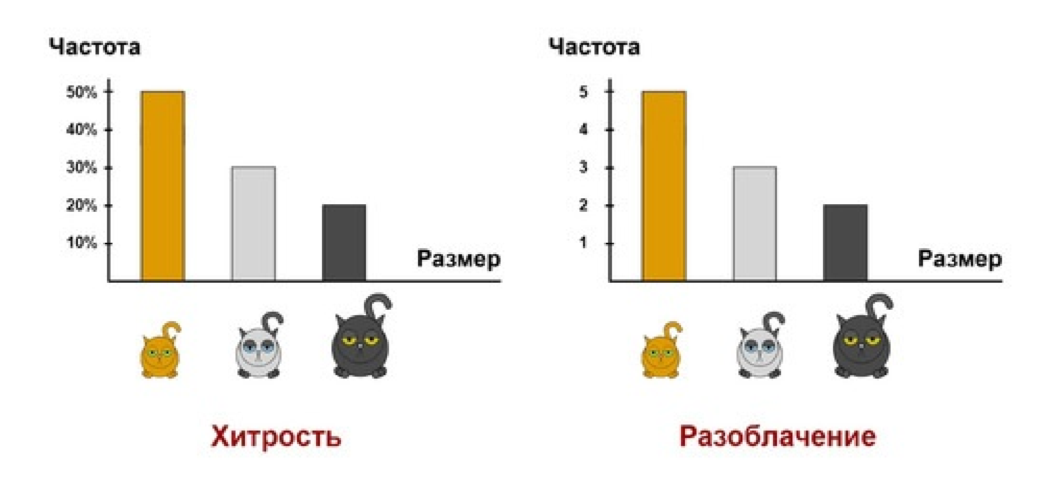
На диаграмме не должно быть ничего лишнего. Если на ней есть элемент, не несущий какой-либо смысловой нагрузки, его лучше убрать. Потому что чем больше лишних элементов, тем менее понятной будет диаграмма.

То же самое касается цветов: лучше ограничить их количество до трех. А если вы готовите графики для публикации, то лучше их вообще делать черно-белыми.

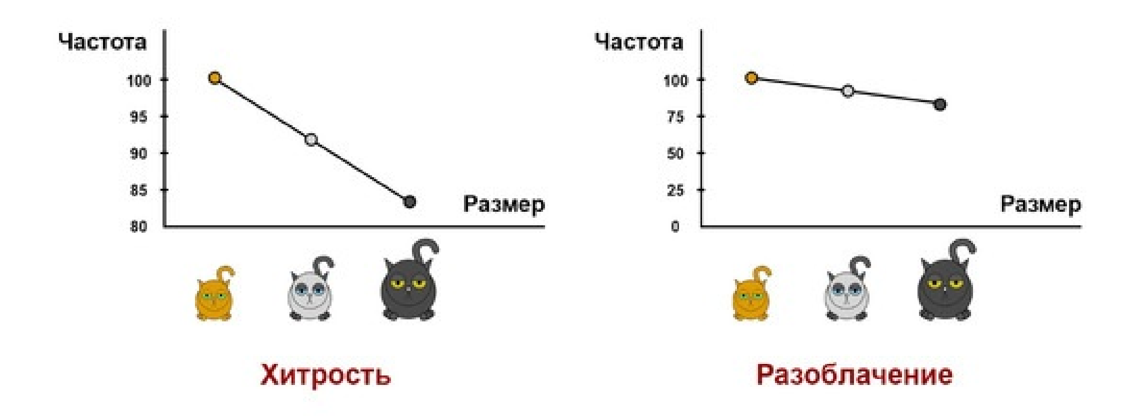
НЕМАЛОВАЖНО ЗНАТЬ!
Темная сторона визуализации

Несмотря на то, что средства визуализации помогают облегчить восприятие данных, они так же легко могут ввести в заблуждение, чем, к сожалению, часто пользуются разные хитрые люди. Ниже мы приведем самые распространенные способы обмана с помощью диаграмм и графиков.

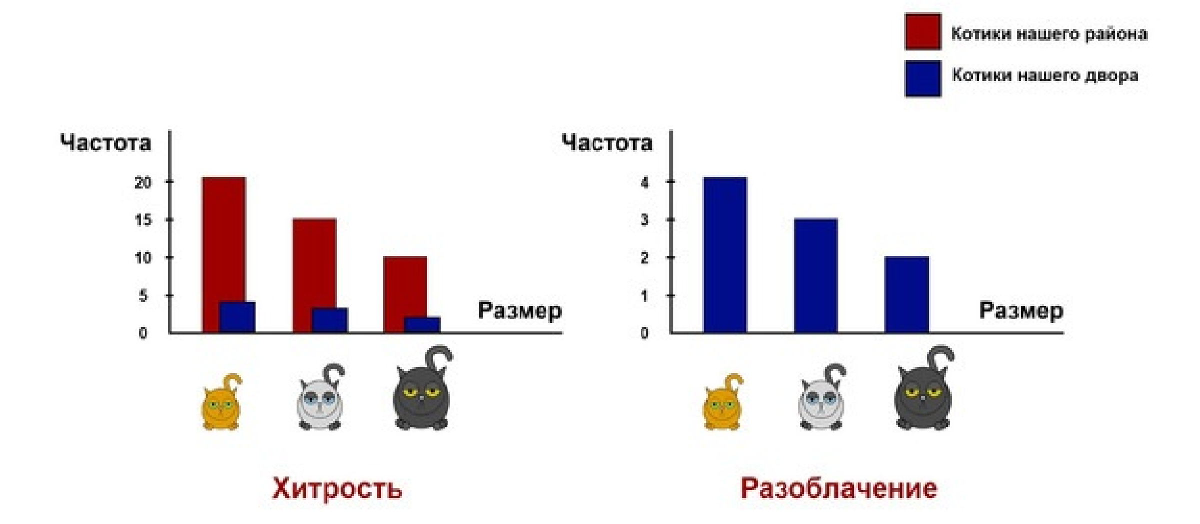
Проценты вместо абсолютных величин. Очень часто, чтобы придать своим данным значимости, хитрые люди переводят абсолютное количество котиков в проценты. Согласитесь, что результаты, полученные на 50% котиков, выглядят куда солиднее, чем на пяти.



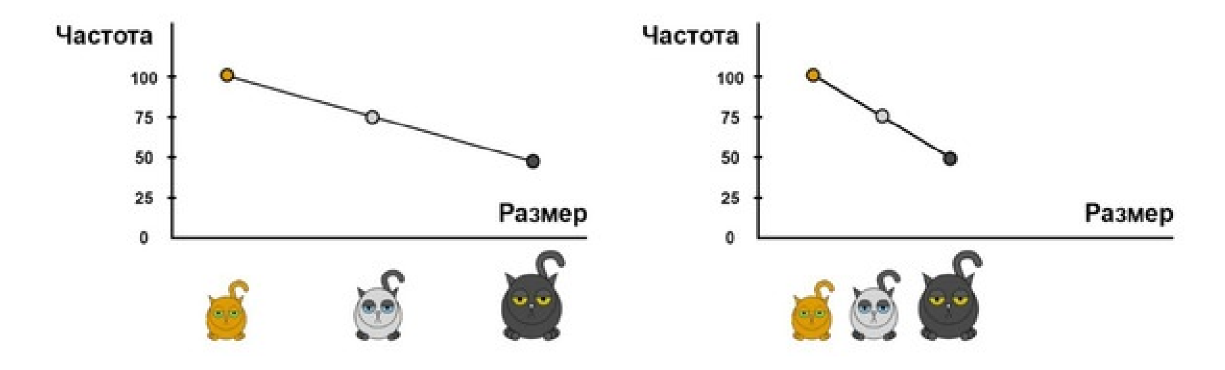
Сдвиг шкалы. Чтобы продемонстрировать значимые различия там, где их нет, хитрые люди как бы «сдвигают» шкалы, начиная отсчет не с нуля, а с более удобного для них числа.



Сокрытие данных. Если же цель хитрого человека в том, чтобы скрыть значимые различия в данных, то их можно разместить на одной шкале с другими данными, которые на порядок отличаются от первых. На их фоне любые различия или изменения будут выглядеть незначительно.



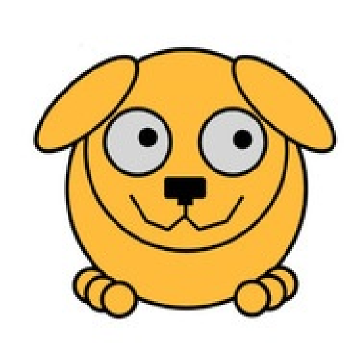
Изменение масштабов. Более мягкий вариант создания иллюзии значимости — это изменение масштабов шкал. В зависимости от масштаба одни и те же данные будут выглядеть по-разному.



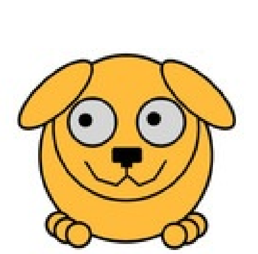
Таким образом, надо быть очень аккуратным, интерпретируя данные, представленные в виде графиков и диаграмм. Гораздо меньше подвержены манипуляции данные, представленные в табличной формуле. Однако и здесь можно использовать некоторые хитрости, которые могут ввести в заблуждение непосвященную публику.

Глава 3.
Чем отличаются котики от песиков или меры различий для несвязанных выборок

Есть котики, а есть песики. Песики чем-то похожи на котиков: у них четыре лапы, хвост и уши. Однако они также во многом различаются — например, котики мяукают, а песики лают.

Но не все различия между ними настолько очевидны. Например, довольно трудно судить о том, различаются ли песики и котики по размеру — ведь есть как очень большие котики, так и очень маленькие песики.




Чтобы понять, насколько они отличаются друг от друга, необходимы так называемые *меры различий для несвязанных выборок*. Большая часть таких мер показывает, насколько типичный песик отличается от типичного котика. Например, самая популярная из них — *t-критерий Стьюдента для несвязанных выборок* — оценивает, насколько различаются их средние размеры.

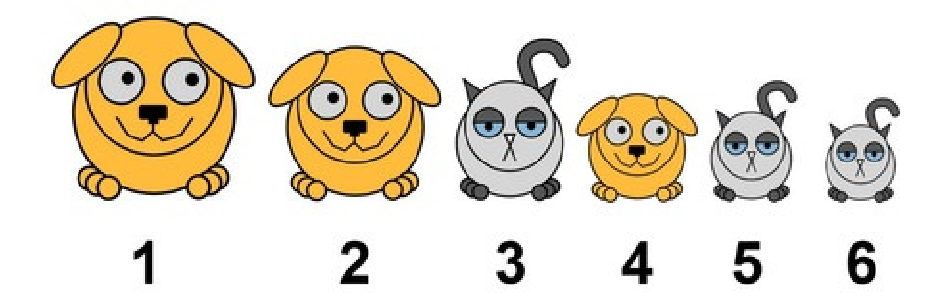
Чтобы рассчитать этот критерий, необходимо из среднего размера песиков вычесть средний размер котиков и поделить их на *стандартную ошибку* этой разности. Последняя вычисляется на основе стандартных отклонений котиковых и песиковых размеров и нужна для приведения t-критерия к нужной размерности.



Если разность средних достаточно большая, а стандартная ошибка очень маленькая, то значение t-критерия будет весьма внушительным. А чем больше t-критерий, тем с большей уверенностью мы можем утверждать, что в среднем песики отличаются от котиков.

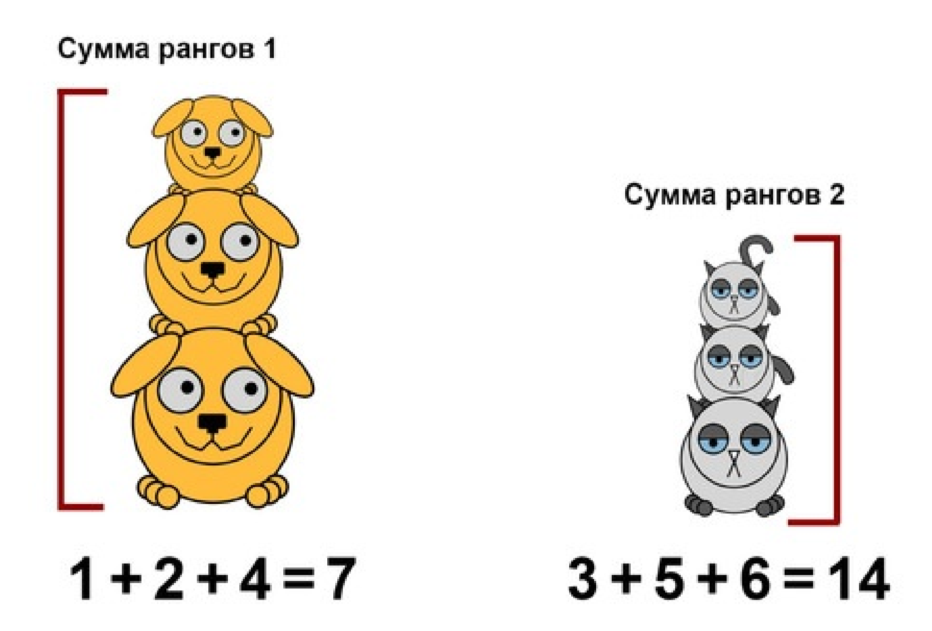
К большому сожалению, поскольку формула t-критерия включает в себя средние значения, то этот критерий будет давать неадекватные результаты при наличии котиков и песиков аномальных размеров (т. е. выбросов, о которых подробно рассказано в первой главе). Чтобы этого избежать, вы можете либо исключить этих котиков и песиков из анализа, либо воспользоваться непараметрическим *U-критерием Манна-Уитни*. Этот критерий, кстати, используется и в тех ситуациях, когда точные (сантиметровые) размеры животных нам неизвестны.

Чтобы рассчитать критерий Манна-Уитни, необходимо выстроить всех песиков и котиков в один ряд, от самого мелкого к самому крупному, и назначить им ранги. Самому большому зверьку достанется первый ранг, а самому маленькому — последний.



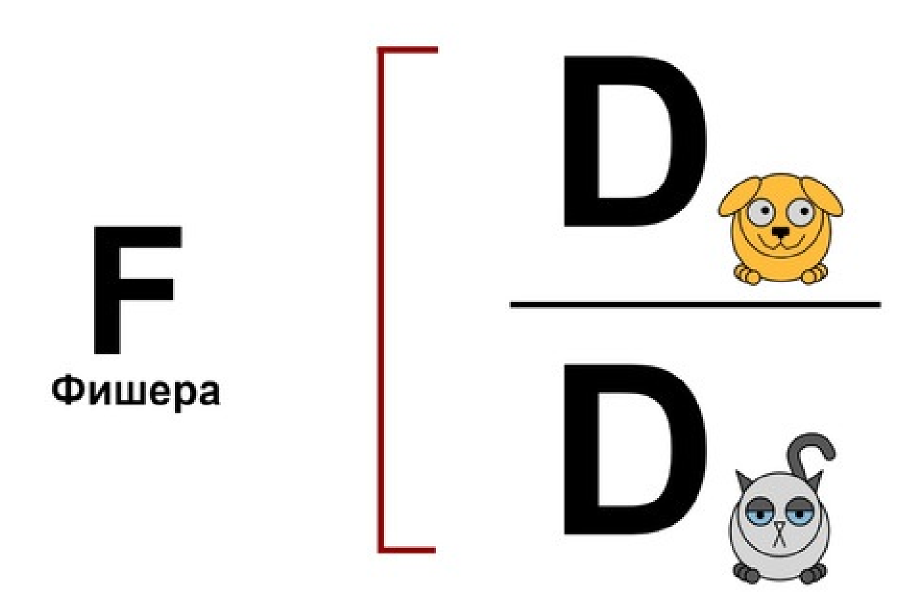
После этого мы снова делим их на две группы и считаем суммы рангов отдельно для

песиков и для котиков. Общая логика такова: чем сильнее будут различаться эти суммы, тем больше различаются песики и котики.



Наконец, мы проводим некоторые преобразования (которые в основном сводятся к поправкам на количество котиков и песиков) и получаем критерий Манна-Уитни, по которому судим, в действительности ли котики и песики отличаются по размеру.

Помимо определения различий между типичными представителями котикового и песикового видов, в некоторых случаях нас могут интересовать различия по их разнообразию. Иными словами, мы можем посмотреть, являются ли песики более разнообразными по размеру, чем котики, или же нет. Для этого мы можем воспользоваться *F-критерием равенства дисперсий Фишера*, который укажет нам, насколько различаются между собой эти показатели.



Необходимо заметить, что в этой формуле сверху всегда должна стоять большая дисперсия, а снизу — меньшая.

Все вышеперечисленные критерии замечательно работают в случаях, когда нам известны точные или хотя бы приблизительные размеры котиков и песиков. Однако такие ситуации

встречаются далеко не всегда. Иногда мы можем иметь только указание на то, является ли наш зверь большим или маленьким. В таких нелегких условиях определить различия между котиками и песиками нам поможет *критерий Хи-квадрат Пирсона*.

Чтобы вычислить этот критерий, нужно построить так называемые *таблицы сопряженности*. В простейшем случае это таблицы 2x2, в каждой ячейке которых — количество (или, по-научному, частота) песиков и котиков определенного размера. Впрочем, бывают таблицы сопряженности и с большим количеством столбцов и строчек.



Очевидно, что если котики и песики как биологические виды не отличаются по размеру, то больших котиков должно быть столько же, сколько и больших песиков (в процентном соотношении). И основная идея критерия Хи-квадрат состоит в том, чтобы сравнить такую таблицу, в которой песики не отличаются от котиков (иначе — *таблицу теоретических частот*), с той, что есть у нас (*таблицей эмпирических частот*).

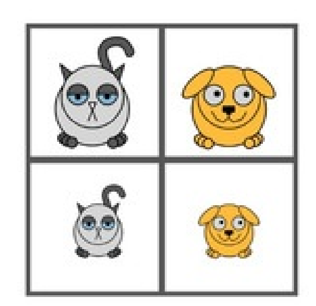


Таблица эмпирических частот

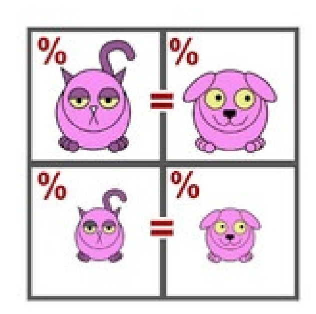
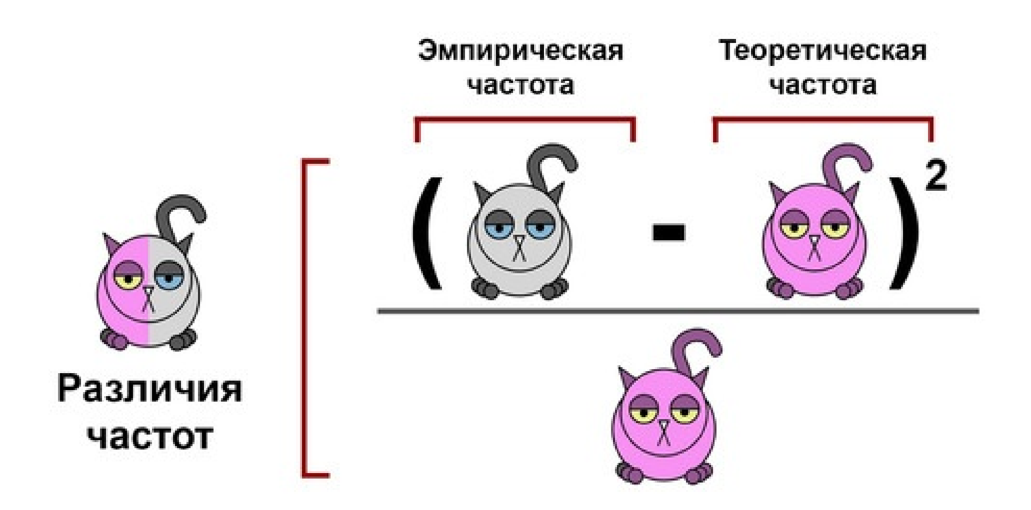


Таблица теоретических частот

Перво-наперво необходимо получить таблицу теоретических частот. Для этого для каждой ячейки подсчитывается *теоретическая частота* по такой формуле.



Следующим шагом мы смотрим, насколько сильно различаются между собой соответствующие ячейки в наших таблицах. Делается это вот так.



Квадрат в числителе этой формулы убирает знак, а знаменатель приводит Хи-квадрат в нужную размерность. Заметим, что если теоретическая частота равна эмпирической, то, применив эту формулу, мы получим 0.

Последним шагом мы складываем все получившиеся значения. Это и будет Хи-квадрат Пирсона. Чем он больше, тем сильнее отличаются песики от котиков.

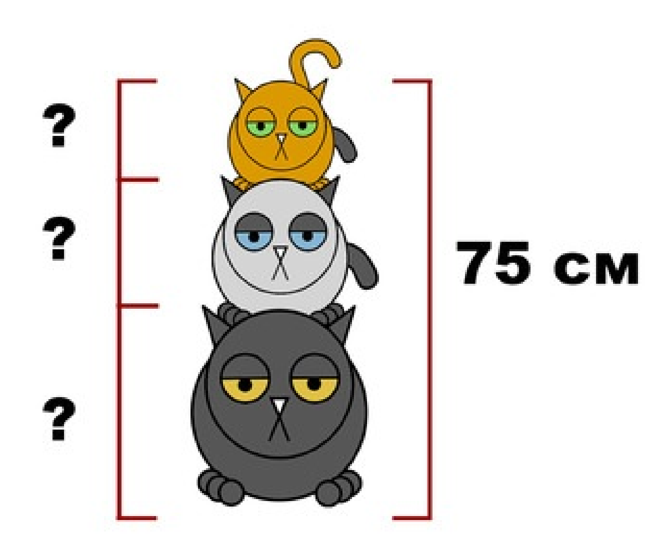


Помимо всего вышеперечисленного существуют и другие статистические критерии, которые позволяют нам определить, чем песики отличаются от котиков. Они, как правило, имеют разные механизмы вычисления и требования к данным. Но вне зависимости от того, каким критерием вы воспользовались, мало просто его вычислить. Необходимо еще и уметь его интерпретировать. И этому вопросу будет посвящена следующая глава.

НЕМАЛОВАЖНО ЗНАТЬ
Загадочные степени свободы

Многих изучающих статистику ставит в тупик понятие «степень свободы», которое часто встречается в учебниках.

Предположим вы знаете, что сумма размеров всех ваших котиков равна 75 см, но не знаете величину каждого конкретного котика. Эти величины будут неизвестны ровно до тех пор, пока вы не начнете их измерять.

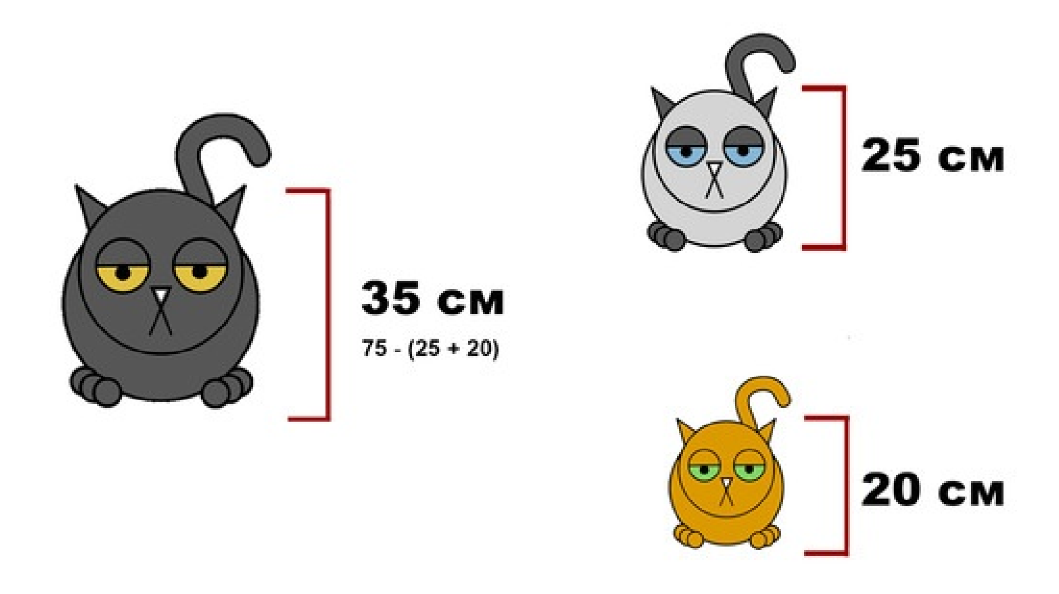


Представим, что вы узнали размер первого котика и он оказался равен 20 см. После

несложных вычислений можно убедиться, что сумма размеров оставшихся котиков будет 55 см. При этом их конкретные размеры до сих пор неизвестны.

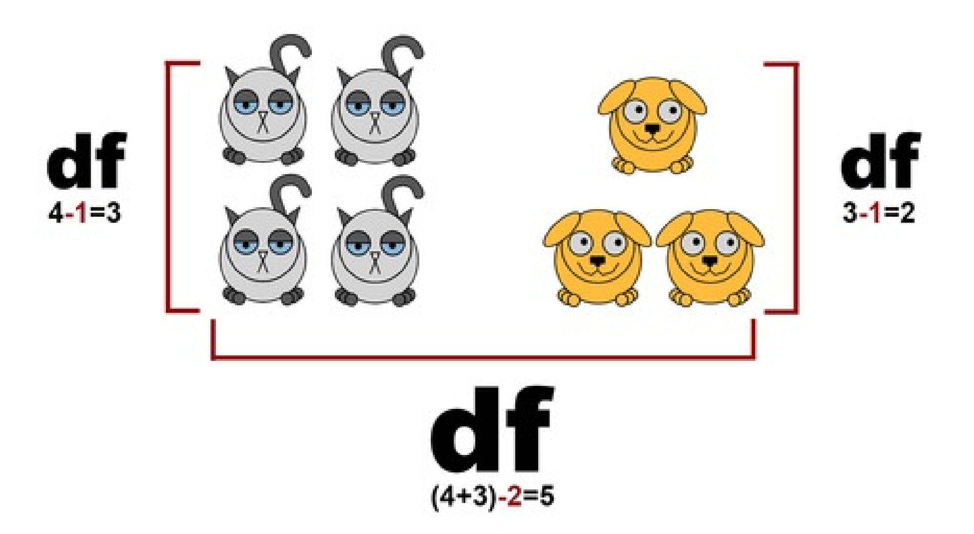


Измерим второго котика. Он оказался равен 25 см. Что мы можем сказать о размере третьего? А то, что он перестал быть неизвестным — теперь мы можем его вычислить. И действительно, вычтя из общей суммы размеры первого и второго котика мы получаем размер третьего.



Число степеней свободы — это то количество котиков, которое мы должны измерить, чтобы однозначно узнать размер всех котиков при известном среднем или дисперсии. Если у вас только одна котиковая выборка, то это количество котиков минус единица.

Если к ним добавляются еще и выборка пёсиков (например, при вычислении t-критерия Стьюдента), то общее количество степеней свободы — это просто сумма степеней свободы котиков и пёсиков. Или по-другому — общее количество животных вычесть двойку.

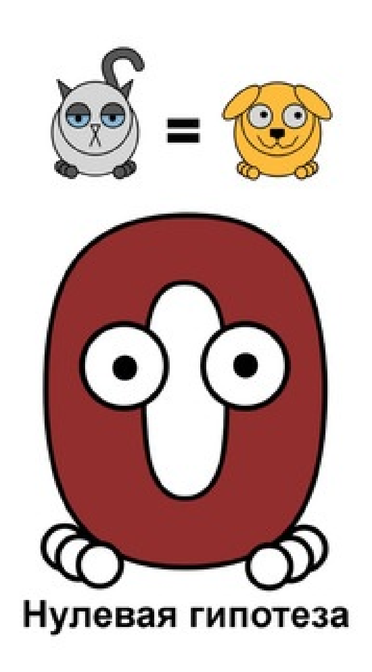


Истоки этого понятия — в самых основах теории вероятности и математической статистики, которые выходят за пределы нашей книги. С практической же точки зрения, знание о степенях свободы нужно при работе с таблицами критических значений и расчёте p-уровня значимости, о которых вы узнаете из следующей главы.

Глава 4.
Как понять, что песики отличаются от котиков или p-уровень значимости

Предположим, что вы вычислили t-критерий Стьюдента. Или U-критерий Манна-Уитни. Или какой-нибудь другой. Как же по нему понять, действительно ли песики и котики различаются по размеру? Чтобы это выяснить, статистики используют весьма нетривиальный подход.

Во-первых, они делают предположение, что котики и песики, как биологические, виды абсолютно не отличаются друг от друга. Это предположение называется *нулевой гипотезой*.

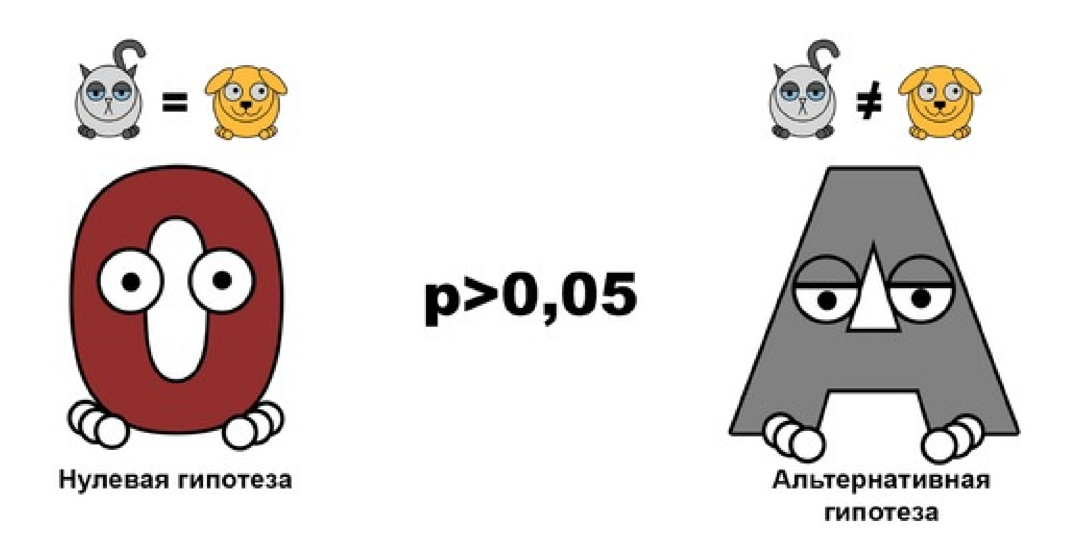


Следующим шагом они вычисляют вероятность того, что две случайно выбранные группы котиков и песиков дадут значение критерия большее или равное тому, которое мы получили (чаще всего без учета его знака). Эта вероятность называется *p-уровнем значимости*.

Если p-уровень значимости меньше 5% (чаще записывается как 0,05), то нулевая гипотеза отвергается и принимается гипотеза о том, что котики и песики все-таки различаются. Такая гипотеза называется *альтернативной*.



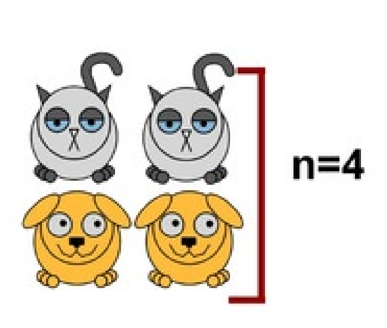
Если же p-уровень значимости больше 0,05, то нулевая гипотеза не отвергается.



Однако то, что она не отвергается, еще не значит, что она верна. Это означает только то, что в данном опыте мы не обнаружили значимых различий.

В специальных статистических программах p-уровень значимости вычисляется автоматически, и нам достаточно просто найти его в соответствующей таблице. Однако, если у вас таких программ нет, то вам придется пользоваться *таблицами критических значений*.

Работать с ними просто: найдите нужную строчку и посмотрите на значение критерия, которое там указано. Если то, что вы получили, превышает это значение, то котики и песики отличаются друг от друга. Правда, для этого правила есть исключения — это U Манна-Уитни и родственные ему критерии.



n-2	Критическое значение для p=0,05
1	12,7
2	4,3
3	3,2

НЕМАЛОВАЖНО ЗНАТЬ!
Альтернативные подходы

Определение различий по p-уровню значимости в последнее время подвергается жесткой критике. Поэтому немаловажно знать о том, что существуют и альтернативные подходы, которые используются при определении значимости полученных результатов.

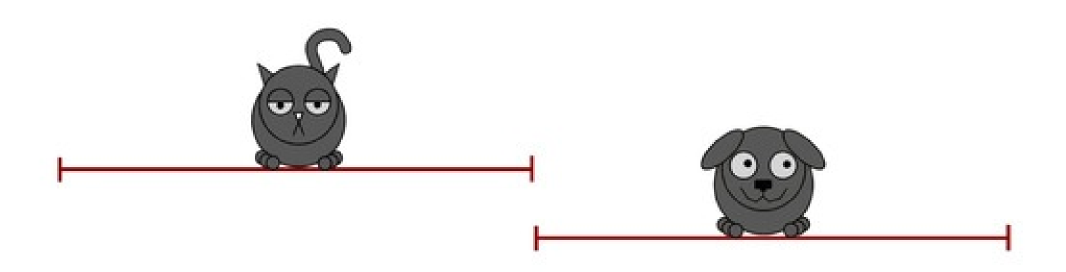
Доверительные интервалы. Как уже было сказано ранее, ученые чаще всего проводят свои исследования не на всех котиках, а на какой-то выборке. Соответственно, они не знают истинного среднего размера по всем котикам. Однако они могут прикинуть, в каком диапазоне он находится. Такой диапазон называется *доверительным интервалом*.



Рядом с доверительным интервалом всегда указывается вероятность. 95%-ый доверительный интервал означает, что мы с точностью в 95% можем утверждать, что истинный средний размер котиков находится в этом диапазоне.

Чем шире такой интервал, тем менее точной считается статистическая оценка. Что касается различий между песиками и котиками, то они имеют место быть, когда их доверительные

интервалы не пересекаются.

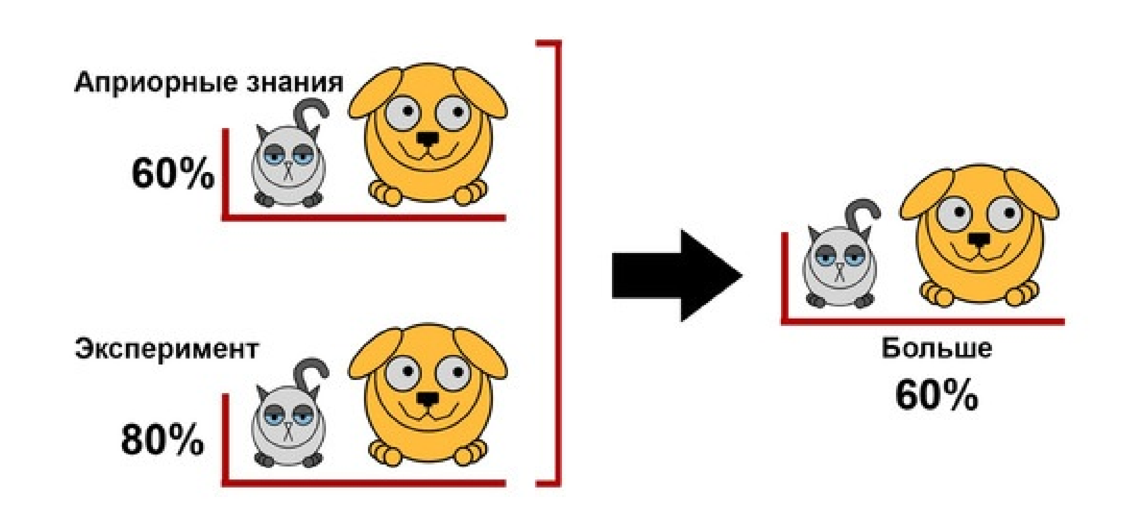


Байесовская статистика. Все вышеприведенные способы определения значимости не учитывают наши предыдущие *(априорные)* знания о том, каких размеров бывают котики и песики. Каждый раз, когда мы определяем p-уровень значимости или доверительный интервал, мы ведем себя так, как будто никогда не видели ни тех, ни других.

Но ведь это не так! Мы ведь достаточно четко представляем себе, как они выглядят! Нельзя просто так брать и отбрасывать предыдущий опыт!

Проблему сопоставления наших предыдущих знаний и новых данных пытается решить группа методов, основанных на теореме английского священника Томаса Байеса.

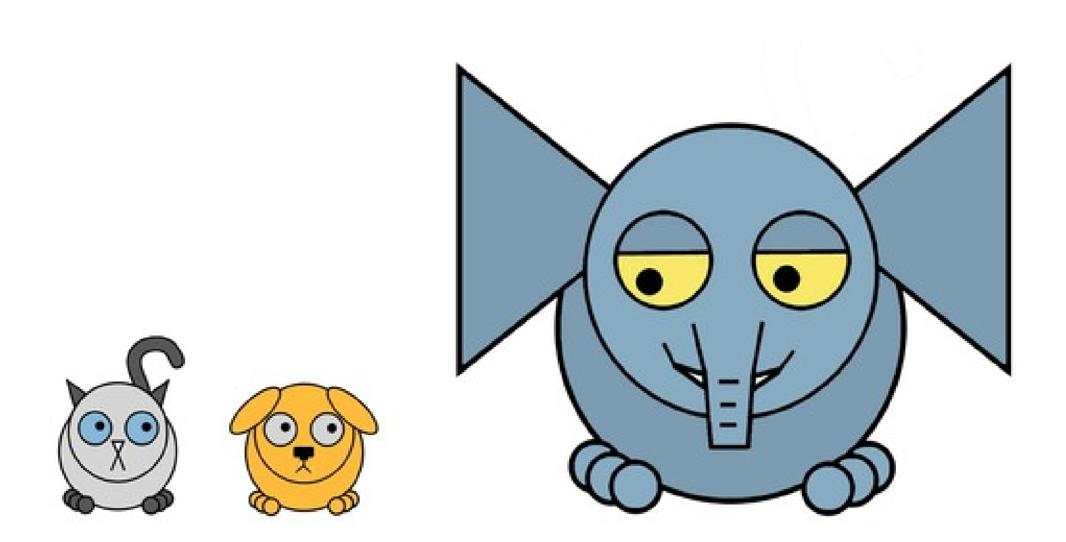
Не вдаваясь в математические подробности, опишем общую логику. Предположим, что из предыдущих опытов мы выяснили, что в 60% случаев случайно выбранный песик больше случайно выбранного котика. Проведя собственный эксперимент, мы обнаружили, что это число гораздо выше — 80%. Следует ли из этого, что нам нужно забыть наш предыдущий опыт и заменить старые данные новыми? Разумеется нет. Новый опыт только подправит предыдущую вероятность, и в следующий раз мы будем считать, что она несколько выше.



Глава 5.
Котики, песики, слоники
или основы дисперсионного анализа

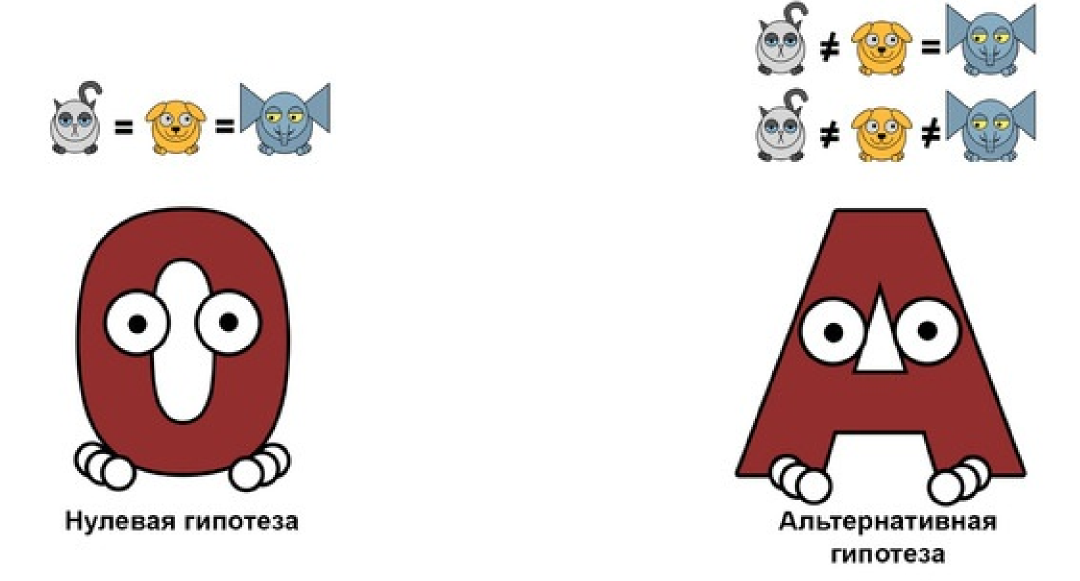
Из предыдущих разделов мы узнали, как определить, различаются ли между собой песики и котики по размеру. И если мы отвечаем на этот вопрос положительно, то мы, по сути, устанавливаем связь между двумя признаками: размером и биологическим видом, к которому принадлежат эти животные.

Однако, согласитесь, что мир не ограничивается только лишь котиками или песиками. Ведь существует еще и множество других животных. Например, слоники.



И, если мы добавим их к нашему небольшому зоопарку, мы не сможем применить обычное попарное сравнение (например, по t-критерию Стьюдента или U-критерию Манна-Уитни) для определения того, связан ли размер с биологическим видом. В этих случаях необходимо использовать другие методы. Например, *дисперсионный анализ*.

Дисперсионный анализ хорош тем, что позволяет сравнивать между собой любое количество групп (две, три, четыре и т. д.) Его нулевая гипотеза состоит в том, что животные абсолютно не различаются между собой по размеру. Альтернативная гипотеза — хотя бы один вид значимо отличается от остальных.



Теперь посмотрим, как это работает.

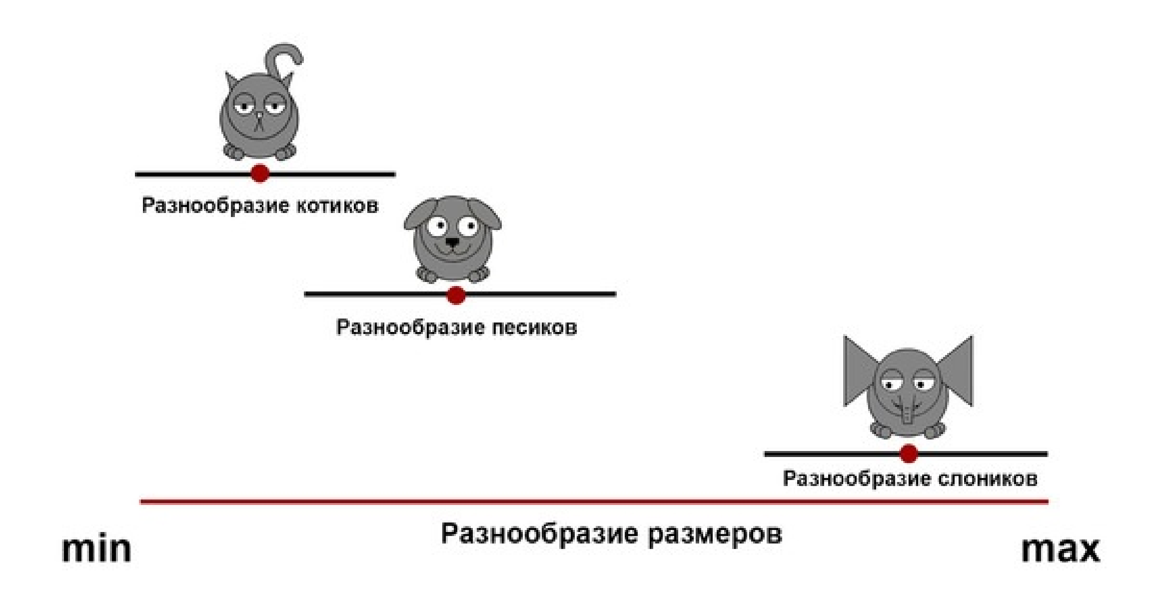
Во-первых, давайте объединим котиков, песиков и слоников вместе и отметим их общее разнообразие. Мы можем заметить, что размеры их типичных представителей могут существенно различаться. Например, средний слоник намного больше среднего котика.



Теперь предположим, что мы убрали отсюда всех слоников. Как вы можете заметить, разнообразие размеров сильно уменьшилось, поскольку слоники вносили в него существенный вклад. И чем сильнее типичные слоники отличались от остальных, тем больше был этот вклад.

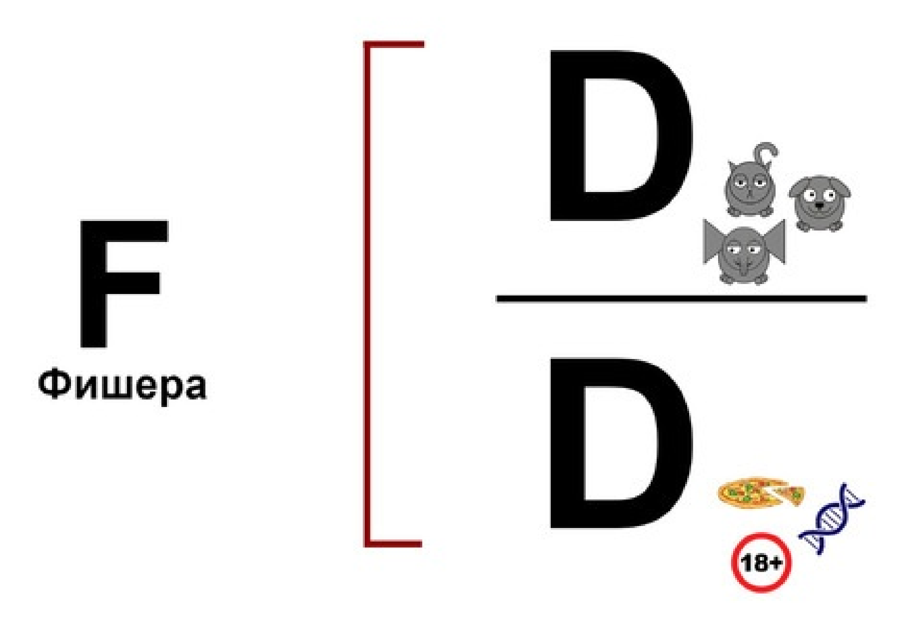


Однако отметим, что котики, песики и слоники по отдельности также бывают весьма различными в зависимости от возраста, генов и режима питания. Теоретически мы можем встретить как очень большого котика, так и весьма маленького слоника.



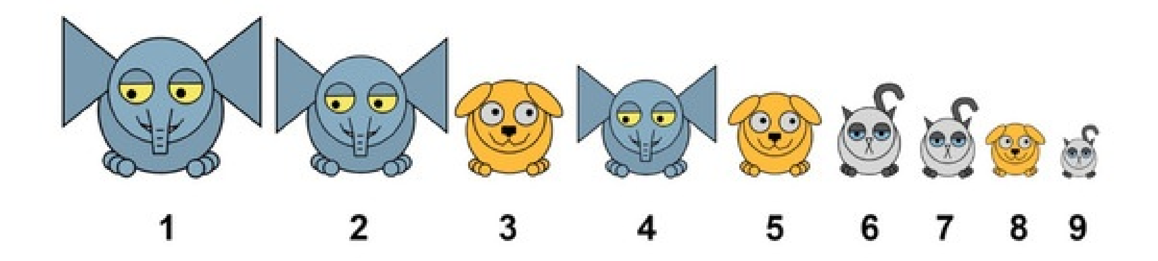
Таким образом, разнообразие размеров складывается как из принадлежности животного к тому или иному виду, так и из абсолютно «левых» факторов. И наша задача — сравнить между собой их вклады.

Как мы помним, одной из основных мер, определяющих разнообразие, является дисперсия. И дисперсионный анализ работает именно с ней. Он выделяет ту часть дисперсии, которая обусловлена фактором вида (*межгрупповую дисперсию*), и ту, которая определяется прочими факторами (*внутригрупповую дисперсию*), а затем сравнивает их по F-критерию Фишера, с которым мы встречались раньше. И чем больше будет значение этого критерия, тем сильнее фактор вида влияет на размер животных.

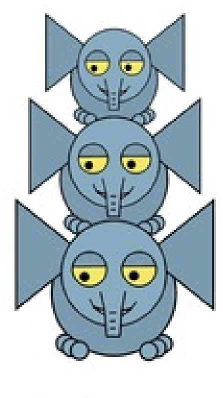


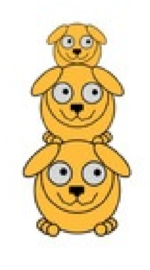
К большому сожалению, дисперсионный анализ является параметрическим методом, а следовательно, не очень любит выбросы и ненормальные распределения данных. Если у вас такая ситуация, то рекомендуется воспользоваться его непараметрическим кузеном — *H-критерием Краскела-Уоллеса*. Последний очень похож на критерий Манна-Уитни, который мы рассматривали в одном из предыдущих разделов.

Мы точно так же объединяем всех животных в одну группу, упорядочиваем их от самого большого до самого маленького и присваиваем им ранги.



Затем они снова делятся на группы, ранги внутри групп складываются, и их суммы сравниваются между собой. Логика здесь такая: чем сильнее различаются суммы рангов, тем больше вероятность отвергнуть нулевую гипотезу. И коэффициент Краскела-Уоллеса как раз и отражает различия в этих суммах.







В заключение напомним, что после вычисления любого из этих критериев необходимо найти соответствующий им p-уровень значимости. Именно он и покажет, существует ли связь между размерами и биологическим видом.

НЕМАЛОВАЖНО ЗНАТЬ!
Проблема множественных сравнений

К большому сожалению, если мы получили значимые результаты по дисперсионному анализу, мы не сможем по ним сказать, кто от кого отличается по размеру: слоники от котиков или песики от слоников. Мало того — мы не можем просто взять и сравнить их попарно с помощью t-критерия Стьюдента. Истоки этого — в основах теории вероятности, и мы не будем на них подробно останавливаться. Просто отметим, что с каждым таким сравнением вы серьезно увеличиваете свои шансы ошибиться в выводах. Эта неприятная вещь называется *проблемой множественных сравнений*.

Поэтому такие сравнения необходимо проводить с помощью других, так называемых *апостериорных критериев* (или критериев *post hoc*).

Простейший из них называется *t-критерием Стьюдента с поправкой Бонферрони*. Вычисляется он как самый обычный t Стьюдента. Поправка же касается критического значения, с которым мы сравниваем p-уровень значимости (0,05). Это значение нужно поделить на количество попарных сравнений.

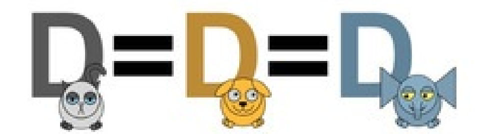


Если вы сравниваете три вида животных, то таких сравнений тоже будет три (котики с песиками, песики со слониками и слоники с котиками). А вот если их четыре, то количество сравнений увеличивается до шести. И тогда критическое значение будет равно 0,05 / 6.

Применив поправку Бонферрони, посмотрите на ваш p-уровень значимости. Если он ниже получившегося значения, то песики и котики различаются, если же нет, то нет.



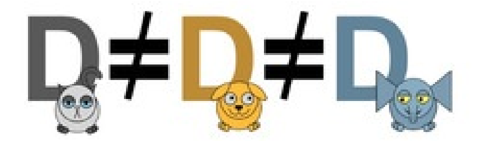
Помимо t-критерия Стьюдента с поправкой Бонферрони существует еще, по крайней мере, 17 апостериорных критериев, которые применяются в различных ситуациях. В первом приближении мы можем разбить их на две группы. В первую входят те критерии, которые применяются, если дисперсии котиков, песиков и слоников не отличаются друг от друга, а вот вторая группа содержит критерии для случая неравных дисперсий. Самые популярные из них представлены ниже.



Поправка Бонферрони

Критерий Шеффе

Критерий Тьюки



Критерий Тамхейна

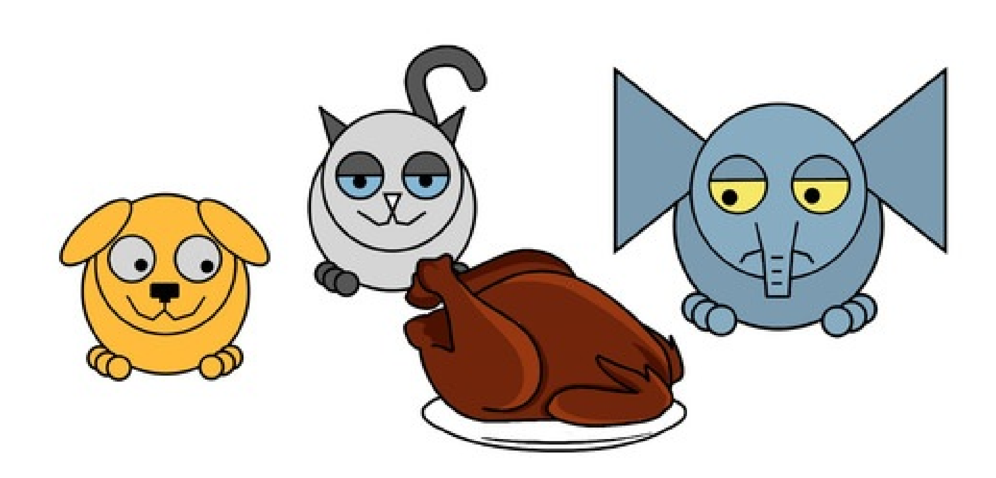
C-критерий Даннета

Критерий Геймса-Хоуэлла

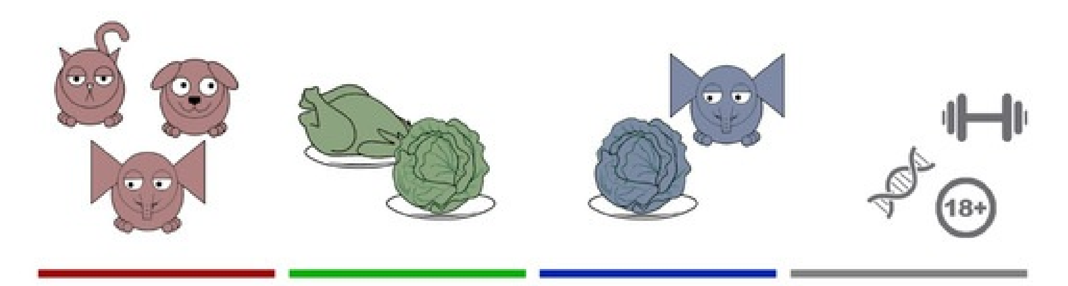
Глава 6.
Диета для котиков
или многофакторный дисперсионный анализ

Из предыдущей главы мы узнали, как определить взаимосвязь между биологическим видом животного и его размером с помощью дисперсионного анализа. Однако, помимо вида, на размер могут повлиять и другие факторы, например, питание.

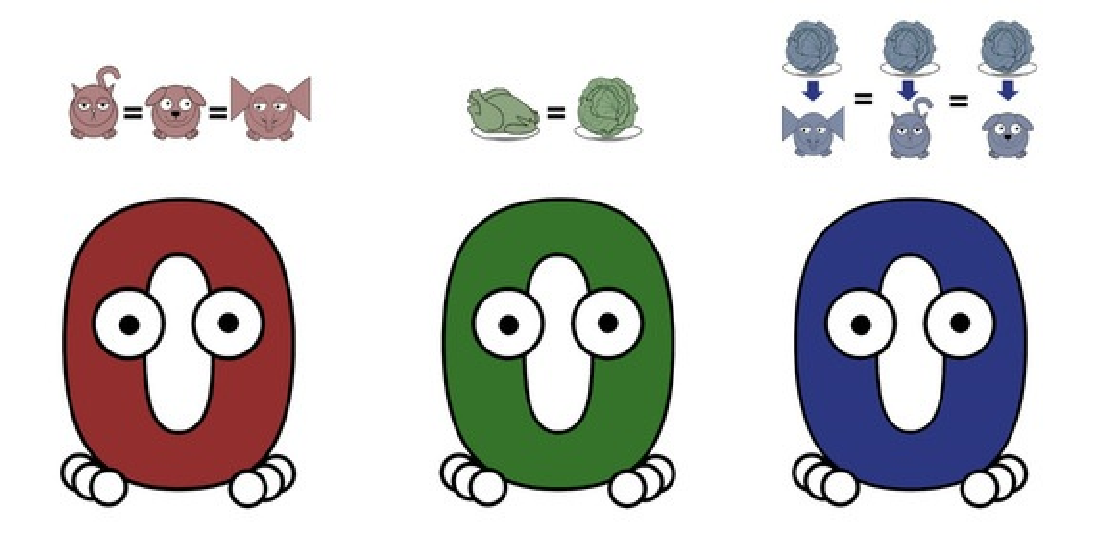
При этом на котиков, песиков и слоников оно может влиять по-разному. Так, мясная диета будет очень нравиться котикам и песикам, в то время как слоники от нее загрустят и будут голодать.



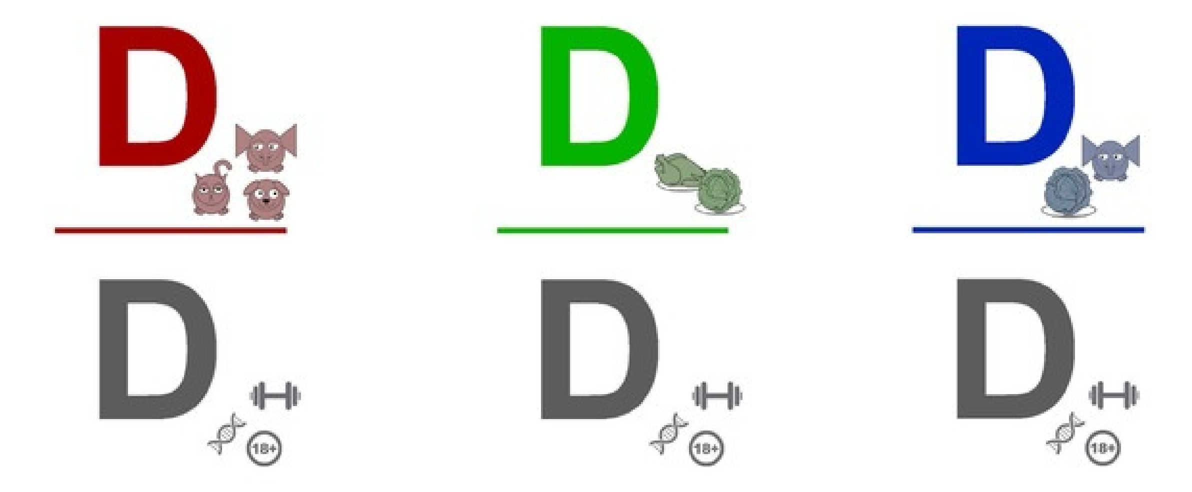
Чтобы разобраться во всех этих влияниях, статистики пользуются *многофакторным дисперсионным анализом*. Простейший из них — *двухфакторный* — разбивает дисперсию на четыре части. Первая отвечает за влияние вида на размер, вторая — за влияние диеты, третья — за взаимодействие этих факторов, а последняя определяется всякими левыми причинами.



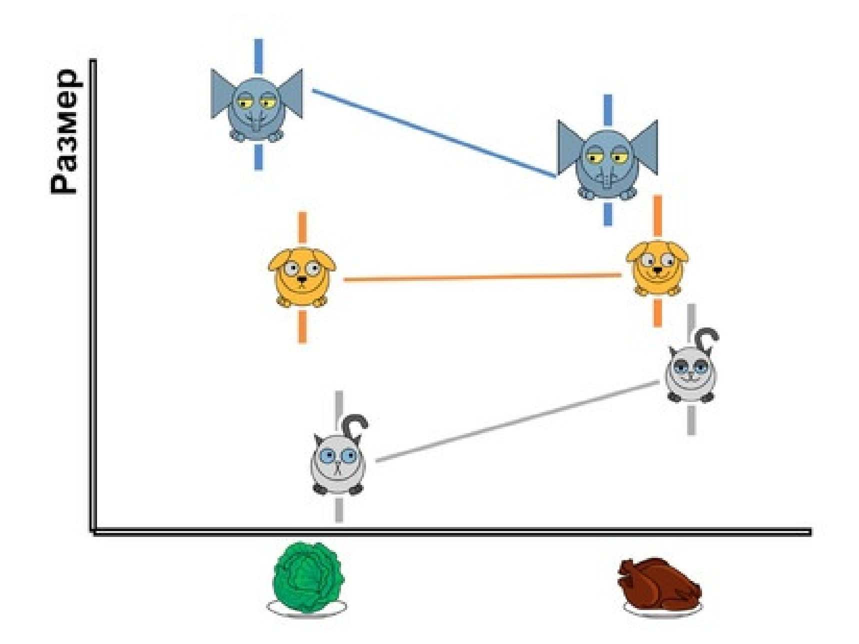
Проверяем мы аж три нулевые гипотезы.
1. Биологический вид не связан с размером.
2. Диета не связана с размером.
3. Диета действует на всех животных одинаково.



Соответственно, для каждой из них считается свой критерий Фишера. И — как и в однофакторном дисперсионном анализе — чем его значение больше, тем больше влияние того или иного фактора.



Для интерпретации результатов двухфакторного дисперсионного анализа легче всего воспользоваться вот такими графиками. Они отражают и средние значения, и дисперсию, и влияние каждого фактора, и их взаимодействие.



В частности из этого графика мы можем сделать следующие выводы.

1. В среднем самые большие животные — слоники, а самые маленькие — котики.

2. Диета по-разному влияет на животных в зависимости от вида. Котики, будучи облигатными хищниками, лучше растут при мясной диете, слоники — наоборот, а вот песикам по большому счету все равно, что есть.

3. Если не учитывать влияние вида, то разные формы диеты не влияют на средний размер животных. Если бы такое влияние существовало, то и котики, и песики, и слоники вырастали бы больше при употреблении мяса, чем при употреблении капусты.

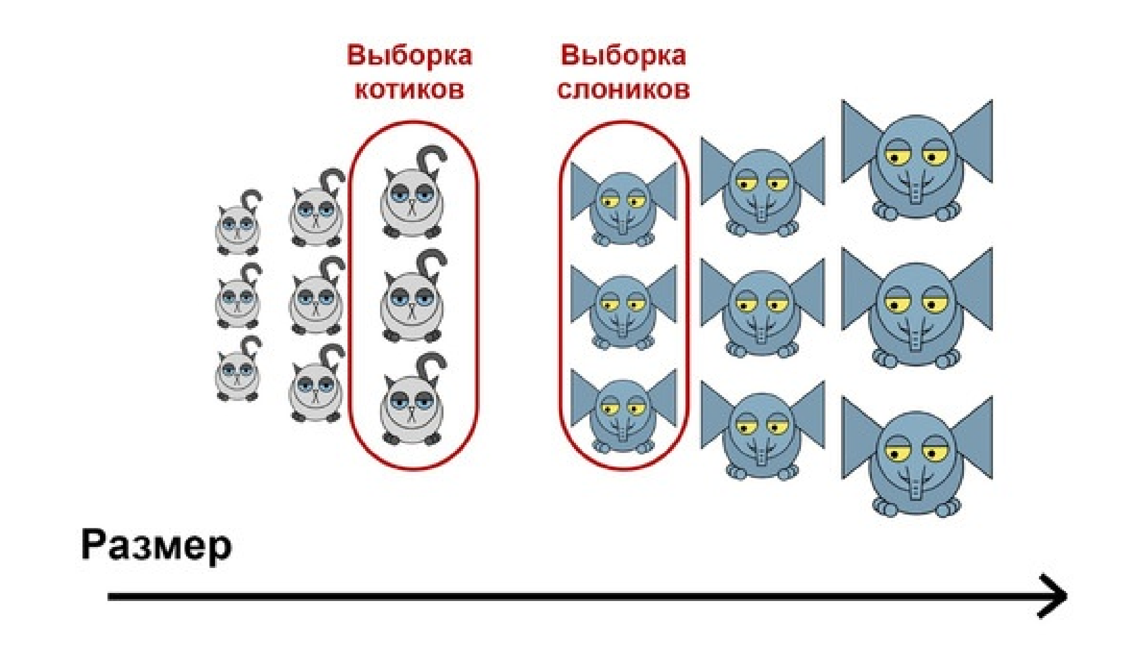
Дисперсионные анализы для трех и более факторов строятся подобным образом: мы проверяем влияние каждого фактора, а также все возможные взаимодействия между ними.

НЕМАЛОВАЖНО ЗНАТЬ!

Сколько нужно котиков?

К настоящему моменту мы продвинулись довольно-таки далеко в вопросах применения статистических критериев для изучения особенностей котиков и других видов животных. Однако за бортом остался очень важный вопрос: сколько котиков необходимо измерить, чтобы критерии давали надежный результат?

Дело в том, что, если вы измерите слишком мало котиков, песиков и слоников, вы можете не зафиксировать даже ощутимые различия. Это может произойти, например, если вам случайно попались очень большие котики и очень маленькие слоники, что при маленьких выборках время от времени случается.



В то же самое время, если вы наберете слишком большую выборку, то даже минимальное отклонение от нулевой гипотезы будет давать значимый результат.

Поэтому котиков должно быть не слишком много и не слишком мало. И чтобы определить, сколько их должно быть, проводятся специальные вычисления.

Оптимальный размер выборки зависит от нескольких факторов, главными из которых являются критический p-уровень значимости (как правило, 0,05 или 0,01) и показатель *мощности критерия*. Последняя определяется как вероятность того, что этот критерий найдет значимые различия там, где они действительно есть. Оптимальным считается показатель мощности в 0,8. Соответственно, в оставшихся 20% случаев критерий пропустит значимые различия.

Оставшиеся факторы определяются самой природой критерия.

В некоторых статистических программах есть специальные калькуляторы мощности. Выбрав необходимый критерий, задав $p < 0{,}05$ и мощность выше 0,8 и проделав некоторые дополнительные операции, вы можете получить количество котиков, необходимое для проведения исследований.

Оптимальное количество котиков

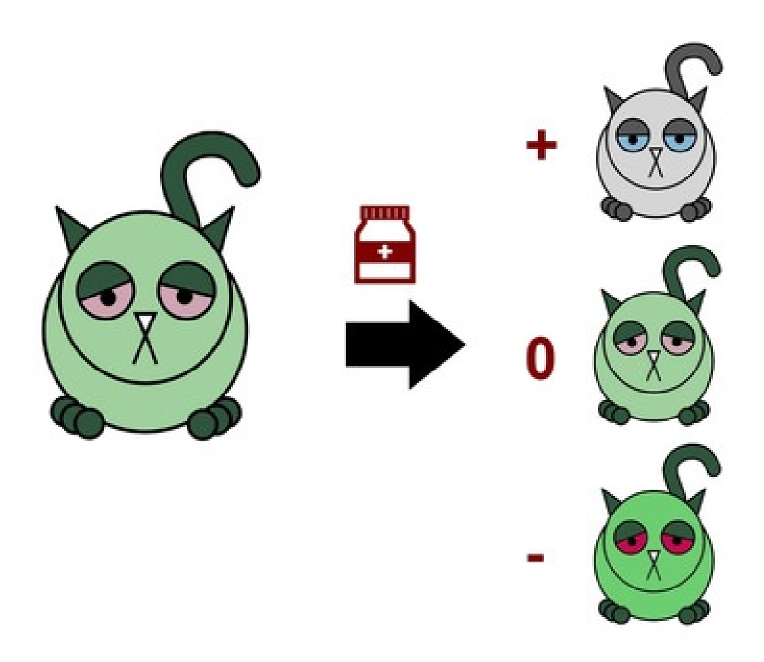
Требуемый p-уровень значимости (0,05)

Требуемая мощность (0,8)

Дополнительные факторы

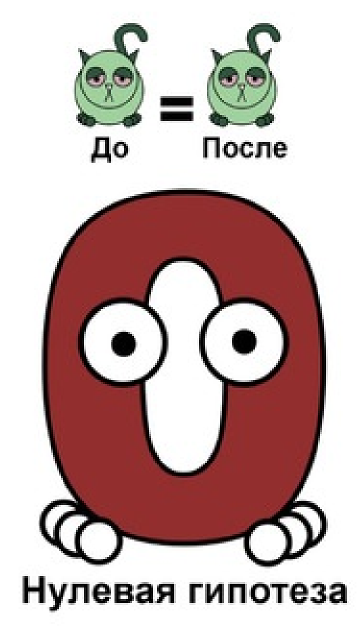
Глава 7.
Что делать, если котик заболел или критерии различий для связанных выборок

Если ваш котик заболел, то его, разумеется, надо лечить. И, как правило, мы делаем это с помощью лекарств. Однако лекарство — штука сложная. Одним котикам оно поможет, на других не повлияет, третьим же может стать хуже.

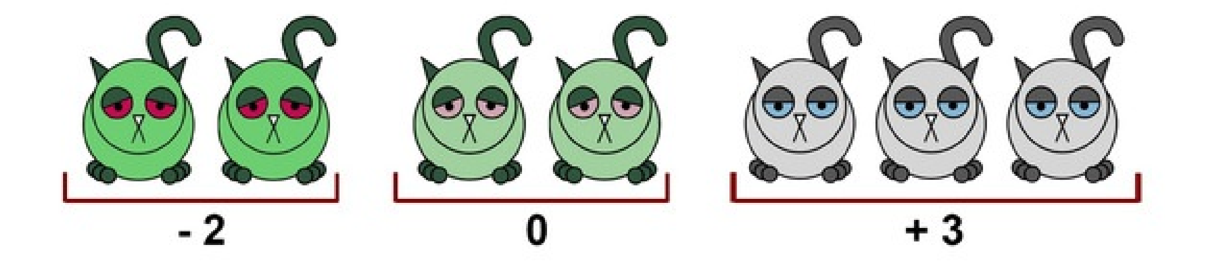


Отсюда вопрос: как понять, можно ли давать лекарство заболевшему котику или нет? Ответ на него могут дать *меры различий для связанных выборок*. Нулевая гипотеза таких критериев — после приема лекарств состояние котиков не изменится.

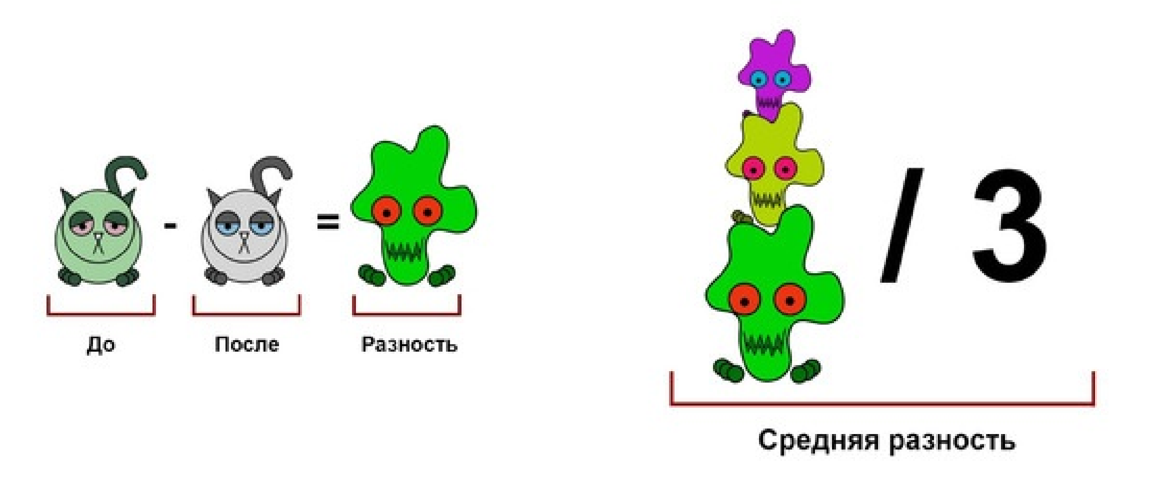
Первое, что приходит в голову, это посчитать количество котиков, которые выздоровели, и число котиков, которым стало хуже, а затем сравнить эти показатели между собой. Котики, на которых лекарство не повлияло, обычно не учитываются.



Такой подход вполне справедлив, и соответствующий метод называется *критерием знаков*. Однако на практике он применяется нечасто, поскольку не позволяет определить, насколько сильно изменилось состояние котиков.

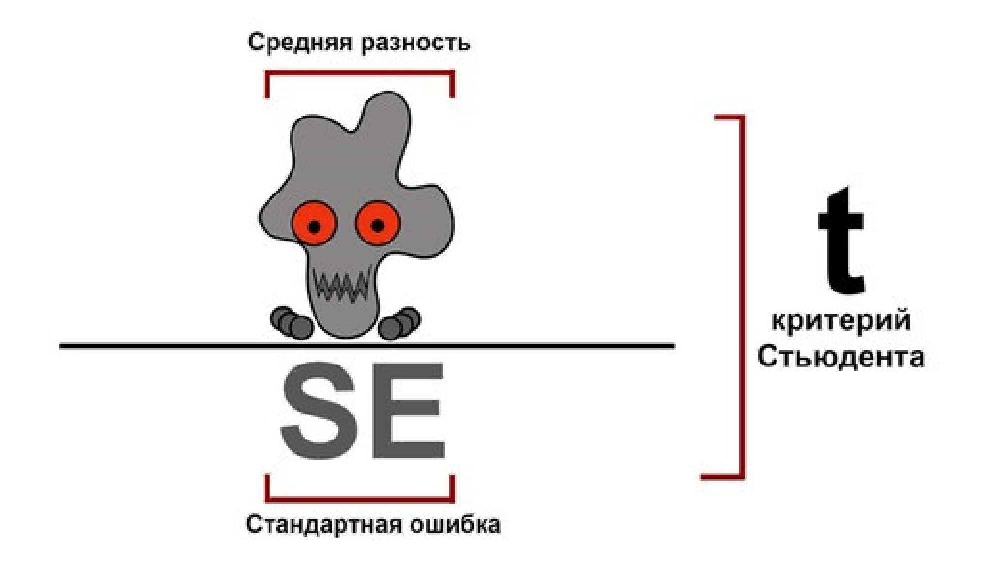


Гораздо чаще мы можем встретить вариант уже известного нам критерия Стьюдента — *t-критерий для связанных (зависимых) выборок*. Идея тут также довольно проста. Сначала мы считаем разности между состоянием каждого котика до и после приема лекарств. Затем мы находим среднее значение от этих разностей.



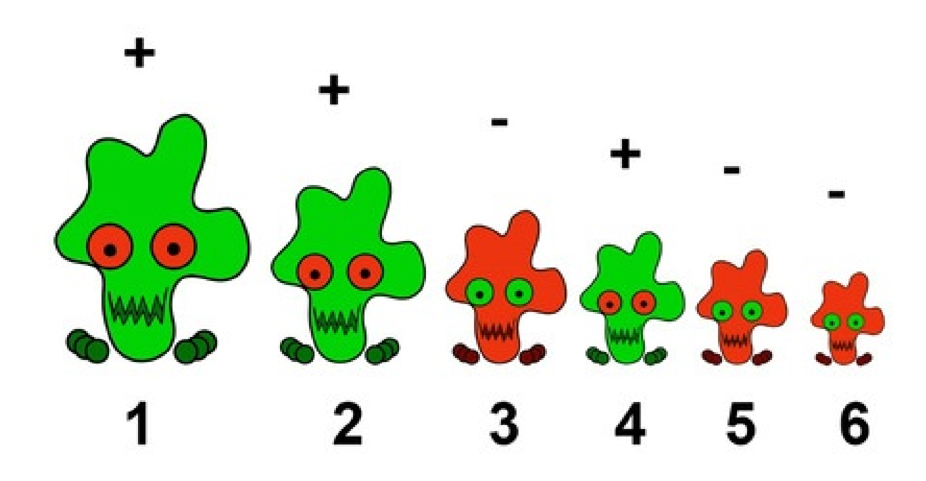
Очевидно, что чем больше это значение, тем сильнее улучшилось или ухудшилось среднее состояние котиков. Если же одной половине котиков стало лучше, а другой — ровно настолько же хуже, то средняя разность будет равна 0.

Завершающим этапом для вычисления t-критерия будет деление средней разности на *стандартную ошибку этой разности*. Как и с обычным критерием Стьюдента, это необходимо для приведения значения к некоторой стандартной размерности. Правда, сама стандартная ошибка считается здесь немного по-другому.

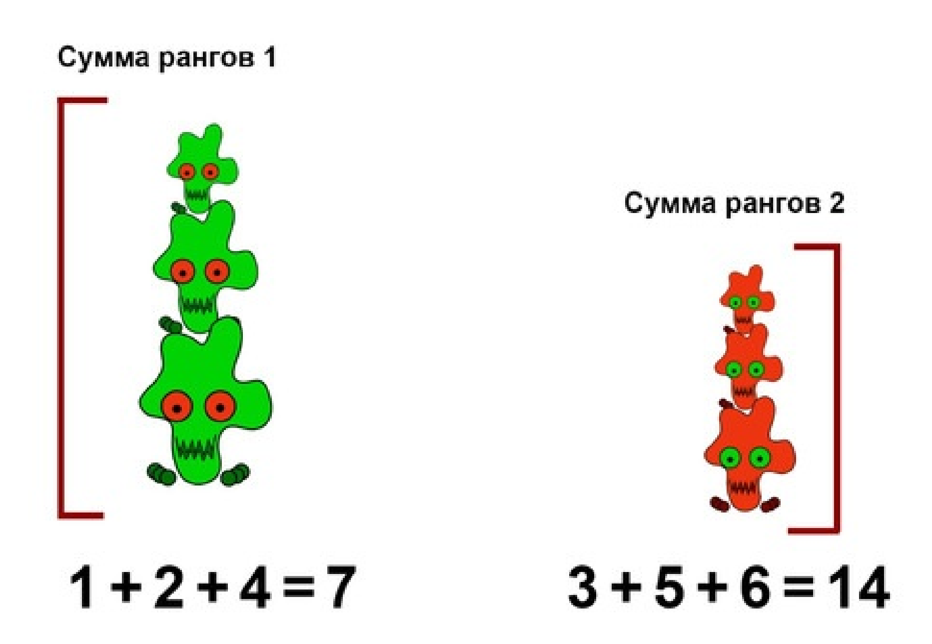


Однако заметим, что, будучи параметрическим (т. е. использующим в своей формуле среднее значение), этот критерий плохо реагирует на выбросы. Поэтому если таковые есть, используйте его непараметрический аналог — *T-критерий Вилкоксона*. Он немного напоминает рассмотренный ранее U-критерий Манна-Уитни.

Итак, чтобы его найти, вычислим разности между состоянием до и после (как и в t-критерии Стьюдента). Затем поставим эти разности в один ряд, от самой большой до самой маленькой, назначив им ранги. При этом знак разности не учитывается.



Теперь снова разделим разности на положительные и отрицательные и посчитаем суммы рангов. Логика здесь такая: чем сильнее суммы рангов будут различаться между собой, тем сильнее улучшается или ухудшается состояние котиков.



Сам Т-критерий можно получить, либо посмотрев на сумму рангов для *нетипичных сдвигов* (т. е. более редких изменений состояния котиков), либо с помощью хитрой формулы, которую мы здесь приводить, пожалуй, не будем.

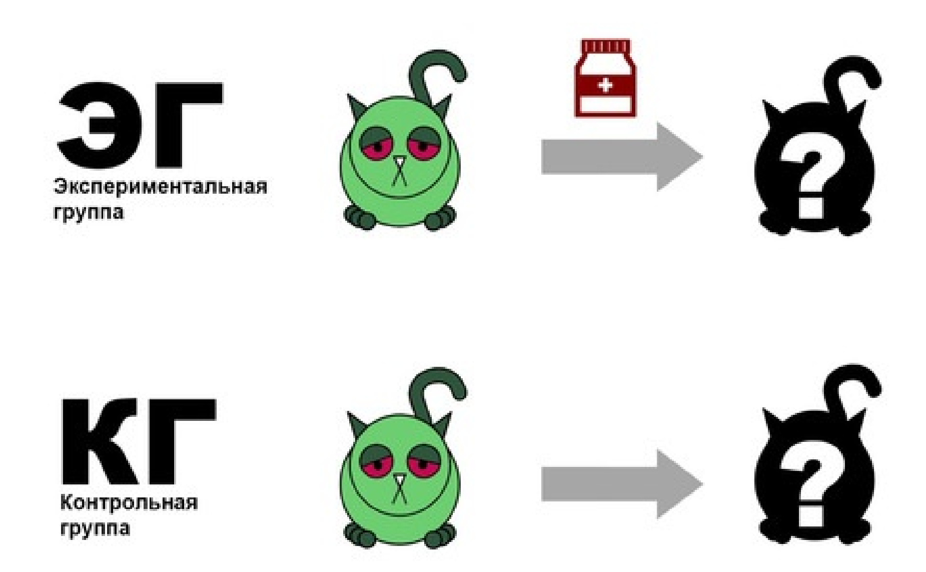
Помимо этих довольно простых методов, для связанных выборок существует свой вариант дисперсионного анализа. Однако о нем мы поговорим уже в следующей главе.

НЕМАЛОВАЖНО ЗНАТЬ!
Эксперимент и как его обработать

Как правило, проверка эффективности того или иного лекарства несколько сложнее, чем описывалось выше. Ведь котики могут выздоравливать и естественным путем. И если мы просто смотрим, как меняется их состояние, то мы не можем быть до конца уверенными, что сильнее

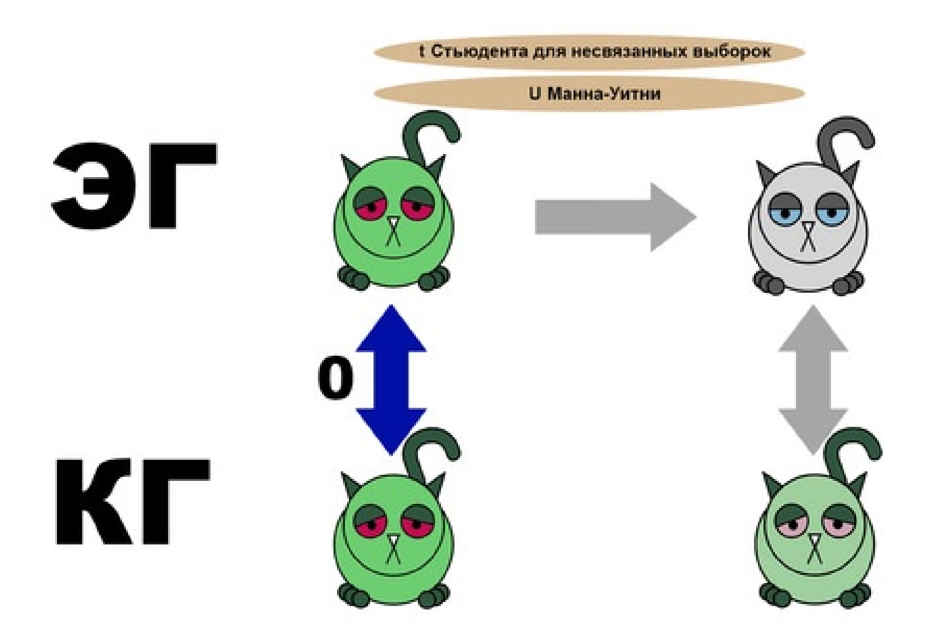
повлияло на них — лекарство или их собственный иммунитет.

Для того чтобы разделить эти влияния, проводят специальную процедуру, называемую *экспериментом*. Для эксперимента требуется две группы котиков — *экспериментальная* и *контрольная*. Первой мы даем лекарство, а вторая лечится своими силами.



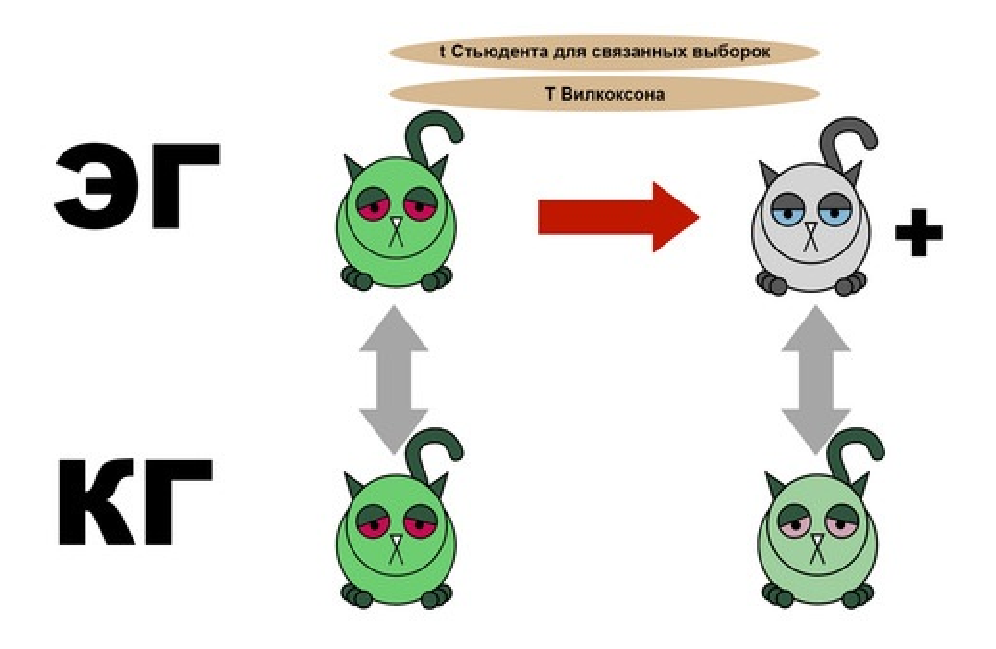
Каждую группу котиков мы замеряем по два раза: до приема и после приема лекарств. Итого мы получаем четыре замера, которые мы сравниваем между собой с помощью мер различий.

Первое, что мы должны сделать, это сравнить группы до эксперимента. Для этого используются t-критерий Стьюдента для несвязанных выборок или U-критерий Манна-Уитни. Котики при этом не должны различаться. Если в одной из групп котики более здоровы, то это очень плохо, поскольку не позволит четко отследить влияние лекарства.

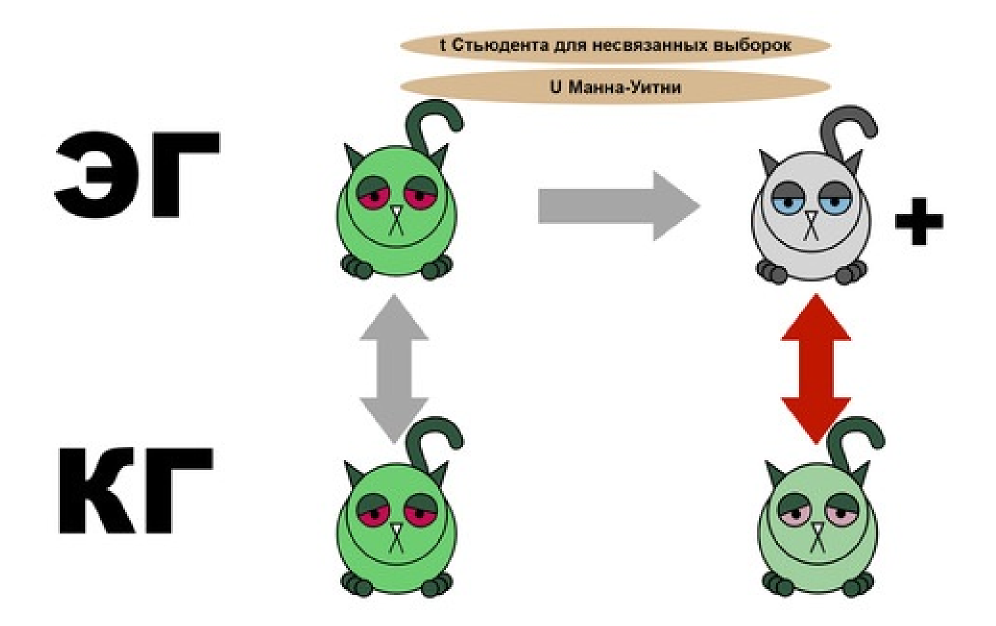


Далее мы сравниваем экспериментальную группу до и после приема лекарств с помощью t Стьюдента для связанных выборок либо T Викоксона. Если различия есть и состояние котиков улучшилось, то мы можем начинать радоваться. Но не сильно. Ведь вполне возможно, что

контрольная группа продемонстрировала тот же результат.

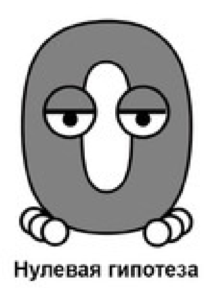


Поэтому последним замером мы смотрим, чем отличаются экспериментальная и контрольная группы после приема лекарств. Если различия есть, и экспериментальным котикам гораздо лучше, чем контрольным, то лекарство реально подействовало.



Таким образом, мы можем сделать вывод, что лекарство действует, только если до эксперимента между группами различий нет, после — есть и имеются положительные изменения состояния в экспериментальной и контрольной группах. Прочие варианты указывают либо на неэффективность лекарства, либо на неправильную организацию эксперимента.

Важно отметить следующее: поскольку для проверки эффективности лекарства мы вычисляли три критерия, то здесь возникает проблема множественных сравнений. Чтобы ее преодолеть, необходимо применить поправку Бонферрони и сравнивать p-уровень значимости не с 0,05, а с 0,017. В противном случае вы рискуете очень сильно ошибиться в своих выводах.



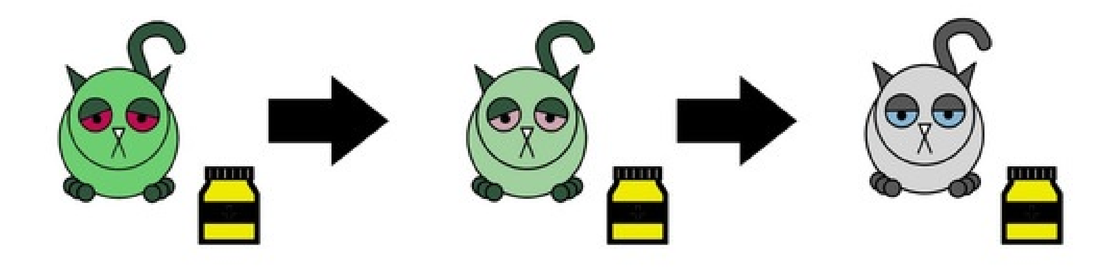
p<0,017



Альтернатива этому — использование *дисперсионного анализа для повторных измерений*, о котором будет рассказано в следующей главе.

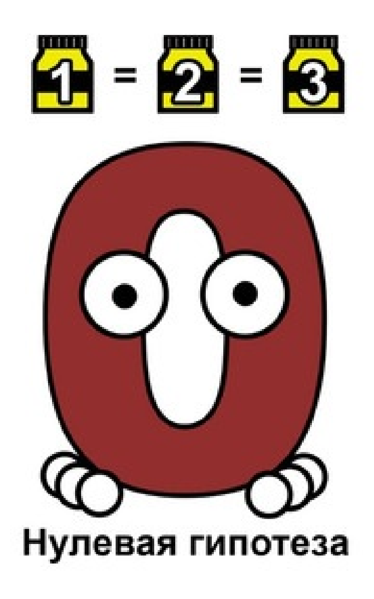
Глава 8.
Лечение котиков
или дисперсионный анализ с повторными измерениями

Из предыдущего раздела мы узнали, как определить, помогает ли то или иное лекарство, если ваш котик заболел. Однако, иногда котики болеют тяжело, и им требуется специальное лечение в особых котиковых клиниках. И, как правило, это лечение подразумевает регулярную сдачу анализов, чтобы отслеживать, становится ли котикам лучше.



Когда таких сдач много (а точнее, больше двух), возникает проблема множественных сравнений, о которой мы не раз говорили выше. Если кратко, то она заключается в том, что, если вы будете попарно сравнивать первый анализ со вторым, второй с третьим и т. д., вероятность того, что вы ошибетесь в своих выводах, будет возрастать.

Разрешить эту проблему, как и в предыдущем случае, может дисперсионный анализ, а точнее, его особая разновидность — *дисперсионный анализ с повторными измерениями*. Нулевая гипотеза такого анализа состоит в том, что состояние котиков от пробы к пробе не меняется.

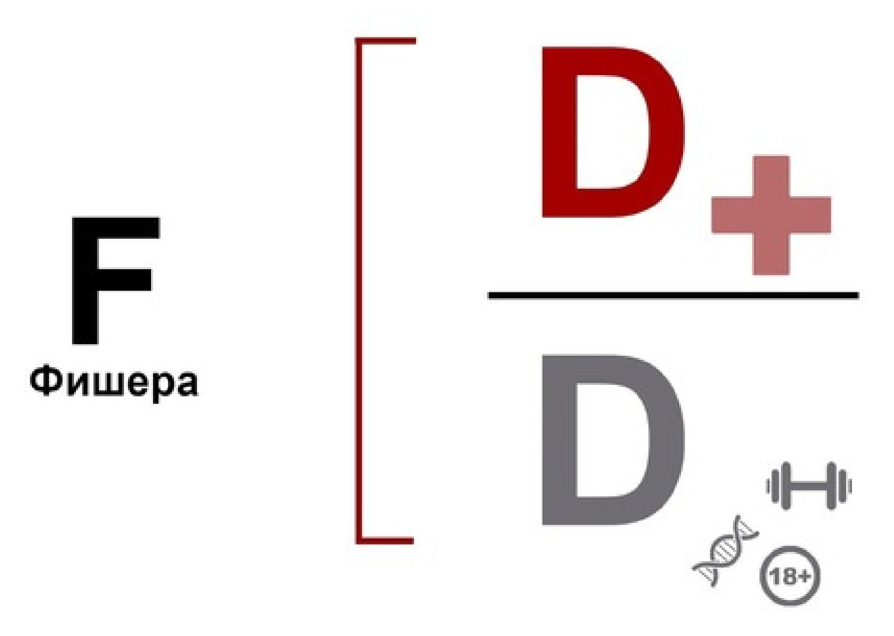


В самом простом варианте мы действуем практически так же, как и при обычном дисперсионном анализе: делим дисперсию на части. В тот раз таких частей было две: первая была обусловлена влиянием лечения (межгрупповая дисперсия), а вторая — остальными факторами (внутригрупповая дисперсия).

Однако важным отличием является то, что мы проводим все измерения на одних и тех же котиках. Иными словами, каждый котик измеряется по несколько раз и, соответственно, вносит свой вклад в общую дисперсию. Таким образом, наша дисперсия делится уже на три части: межгрупповую, внутригрупповую и *межиндивидуальную*.



Критерий Фишера сравнивает между собой только первые два вклада. Соответственно, чем он больше, тем больше причин отклонить нулевую гипотезу. И опять же — если вы отклонили ее, то попарное сравнение нужно будет проводить с помощью специальных post hoc критериев.

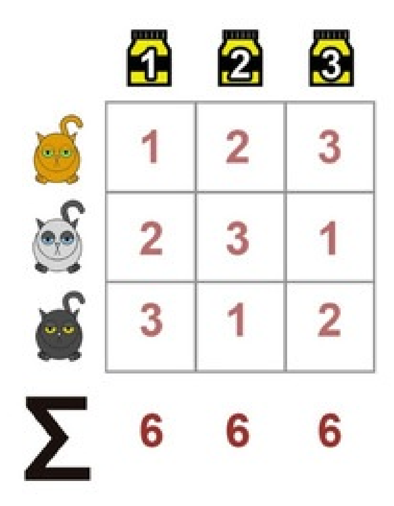


У дисперсионного анализа с повторными измерениями есть свой непараметрический брат-близнец — *критерий Фридмана*, который применяется, если есть выбросы и/или распределение отличается от нормального.

Идея его достаточно проста. Возьмем одного из котиков, у которого взяли три пробы анализов. Каждой из этих проб мы присваиваем ранг, где один — это самый плохой анализ, а три — самый хороший. То же самое мы делаем и с остальными котиками, получая в итоге вот такую таблицу.

	1	2	3
	1	2	3
	1	2	3
	1	2	3
Σ	3	6	9

Очевидно, что если первая проба у всех котиков самая плохая, а последняя — самая хорошая, то по итогу суммы рангов будут сильно различаться и нулевая гипотеза будет опровергнута. Обратная ситуация — когда суммы рангов во всех пробах одинаковы. Это будет означать, что лечение никак не повлияло на котиков.



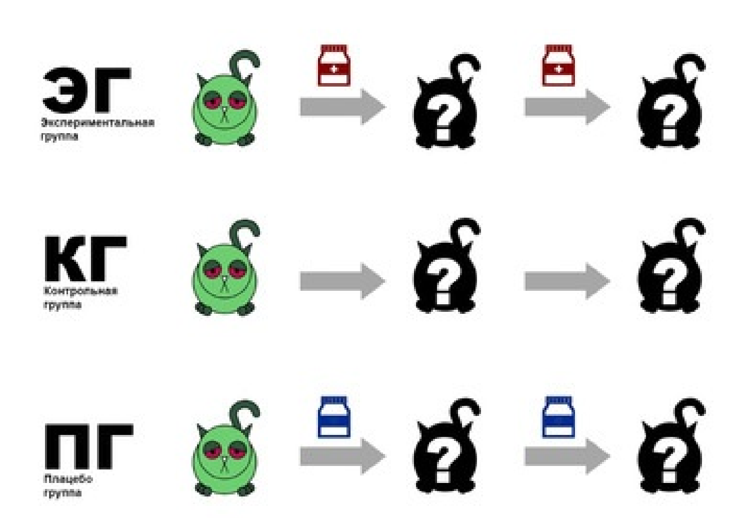
	1	2	3
	1	2	3
	2	3	1
	3	1	2
Σ	6	6	6

Сам же критерий Фридмана, собственно, и позволяет оценить, насколько различаются эти суммы рангов.

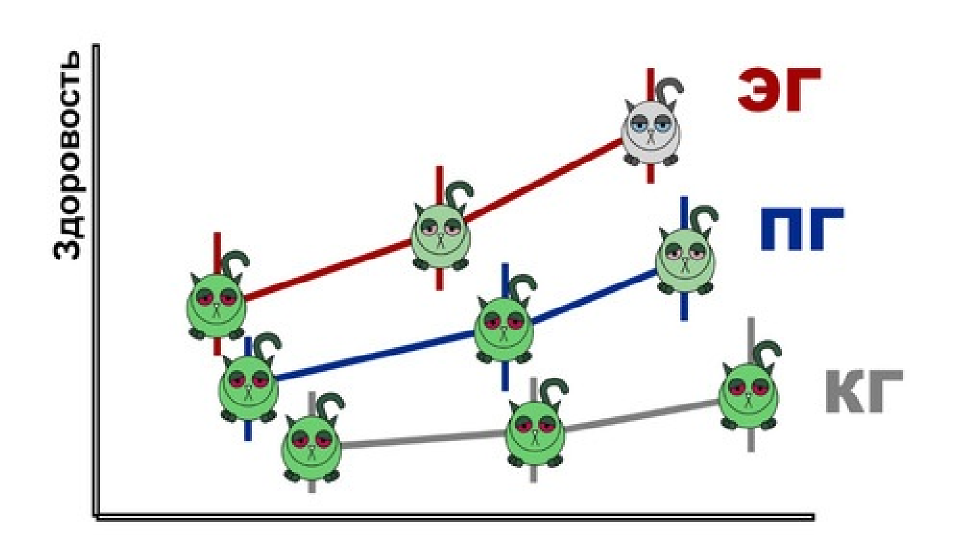
НЕМАЛОВАЖНО ЗНАТЬ!
Сложные эксперименты

Некоторое время назад мы рассмотрели, как правильно обрабатывать простые эксперименты с двумя группами и двумя замерами (до и после воздействия). Однако если групп и замеров больше, то наша задача существенно усложняется.

К примеру, мы разделили наших котиков на три группы: первой мы даем лекарство (экспериментальная), второй не даем лекарство (контрольная), а третьей даем пустышку, но говорим им, что дали лекарство (*плацебо-группа*). При этом каждая группа замеряется три раза: в начале, середине и конце лечения.



Для обработки такого исследования нам необходим двухфакторный дисперсионный анализ с повторными измерениями. Подобно обычному двухфакторному ДА такой анализ легче всего интерпретируется с помощью графиков.

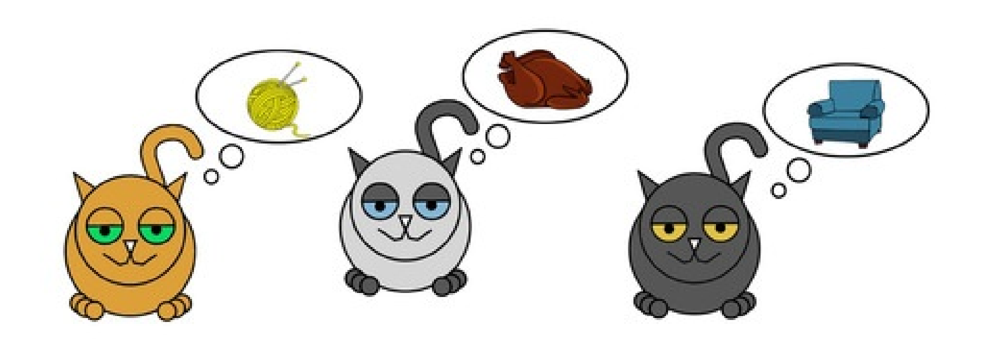


В частности из этого графика мы можем увидеть, что котики, принимавшие лекарство, выздоровели, плацебо-котикам стало чуть лучше, а контрольные котики так и продолжают болеть. Правда, возможно, на наши результаты могли повлиять небольшие различия между котиками в начале эксперимента.

К слову, все попарные различия между группами в разные моменты также необходимо проверять с помощью post hoc критериев. В частности — с помощью поправки Бонферрони.

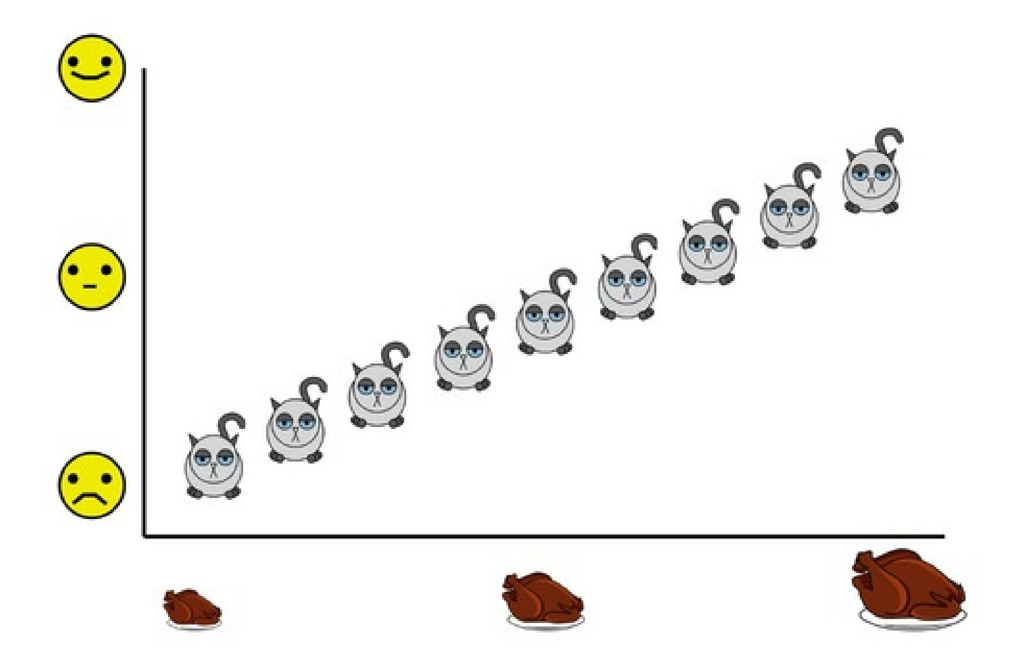
Глава 9.
Как сделать котика счастливым или основы корреляционного анализа

Безусловно, мы все хотим, чтобы наши котики были счастливы, и поэтому стараемся их постоянно радовать. Однако разных котиков радуют разные вещи: один любит вкусно поесть, другой — поиграть, а третий — поточить когти о любимый хозяйский диван.

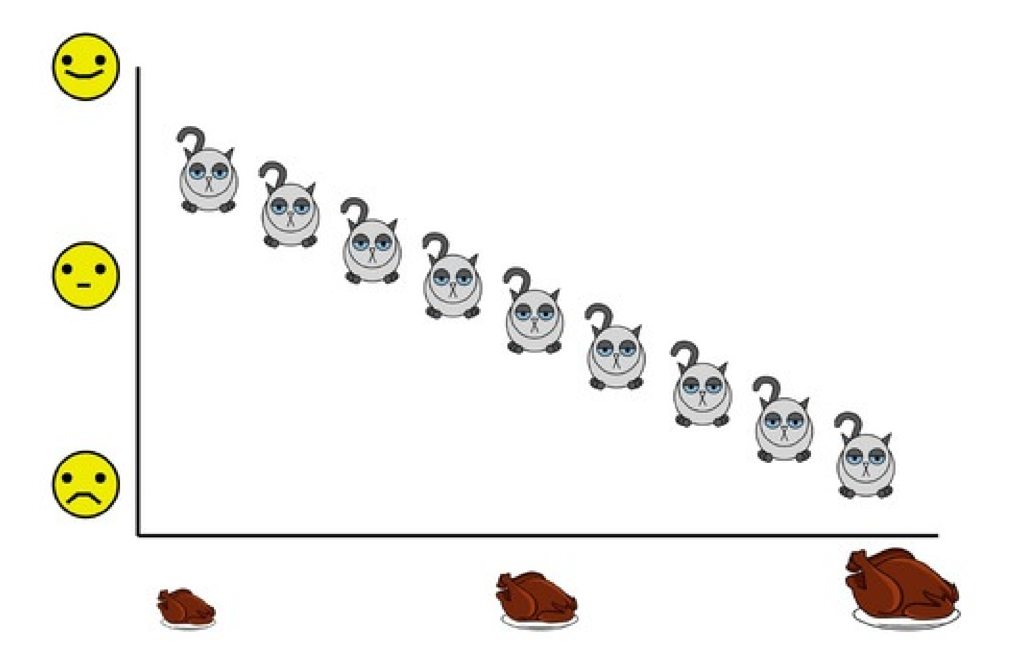


Безусловно, существуют и некоторые универсальные вещи, которые радуют большинство котиков, что сильно упрощает нам жизнь. И в этой главе мы рассмотрим один из методов, который позволяет их выявить, — *корреляционный анализ*.

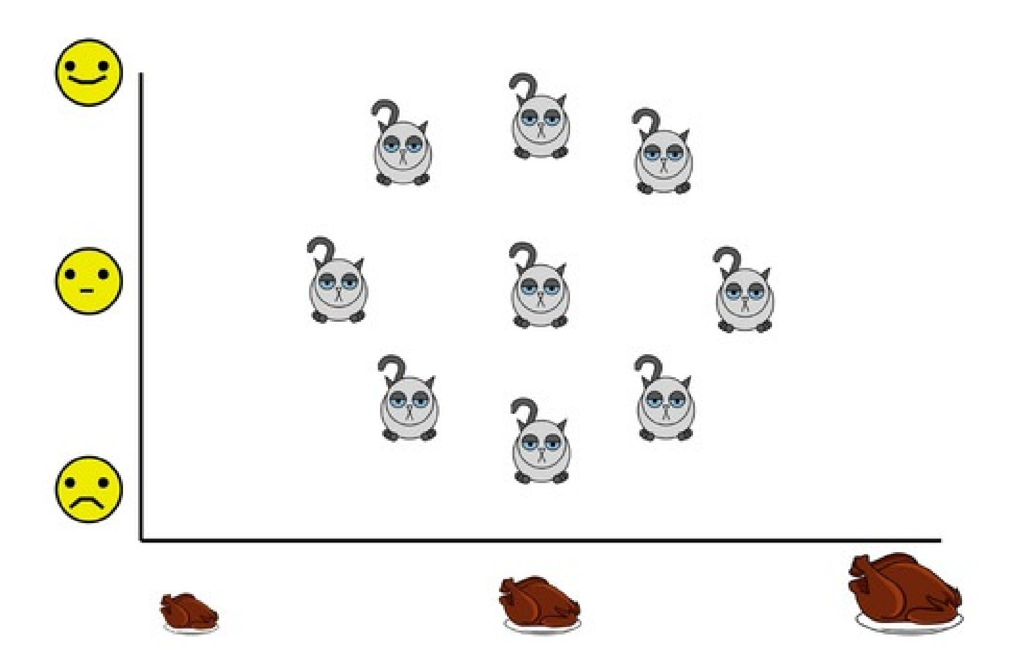
Предположим, мы решили проверить, связаны ли между собой котиковое счастье и размер ежедневных котиковых порций. Если обильная еда делает котиков счастливыми, то эта взаимосвязь будет отражаться вот таким графиком.



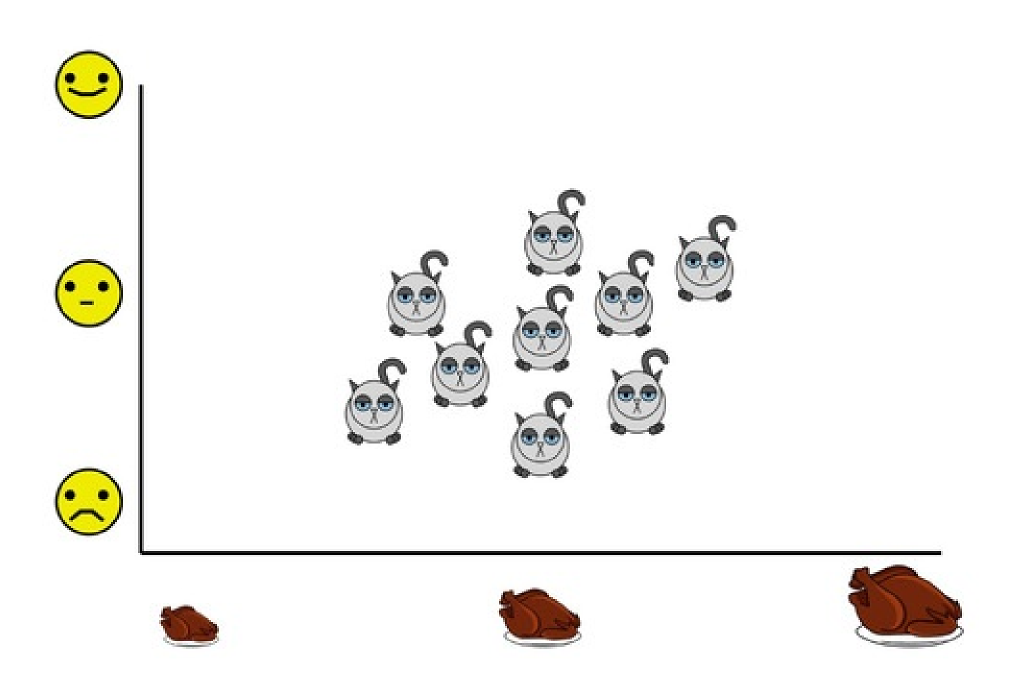
Это так называемая *линейная положительная связь*. Противоположная (хотя и маловероятная) ситуация — котики являются приверженцами оздоровительных голоданий, и чем больше порции им предлагают, тем более несчастными они становятся.



Такая связь называется *линейной отрицательной*. Наконец, может получиться так, что котикам вообще не важно, насколько большие у них порции, главное, чтоб еда была вкусной. В этом случае мы наблюдаем отсутствие связи (или *нулевую связь*), которая отображается вот таким вот графиком.

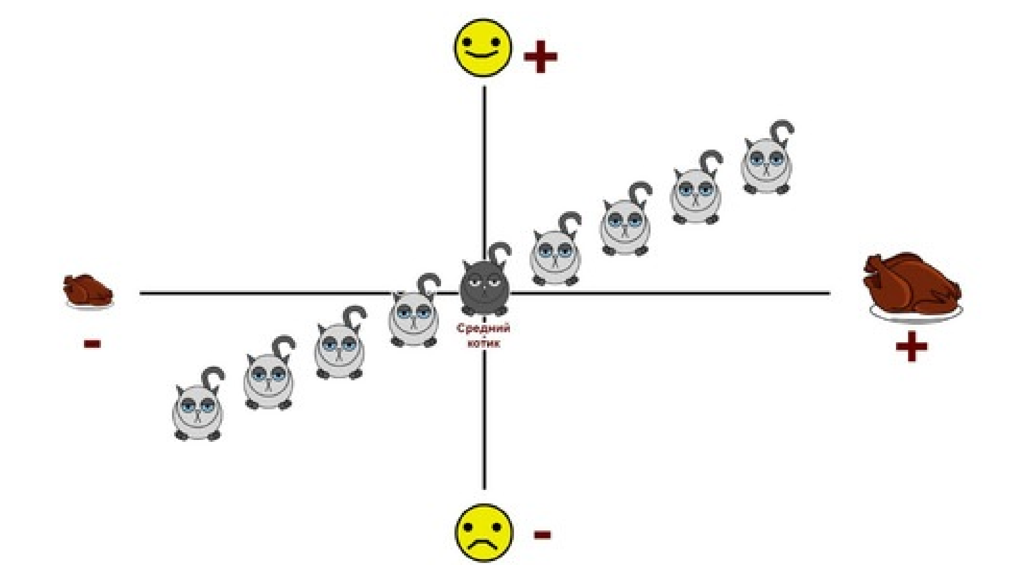


Однако в реальной жизни мы очень редко можем наблюдать подобные случаи: как правило, у нас возникает что-нибудь такое.

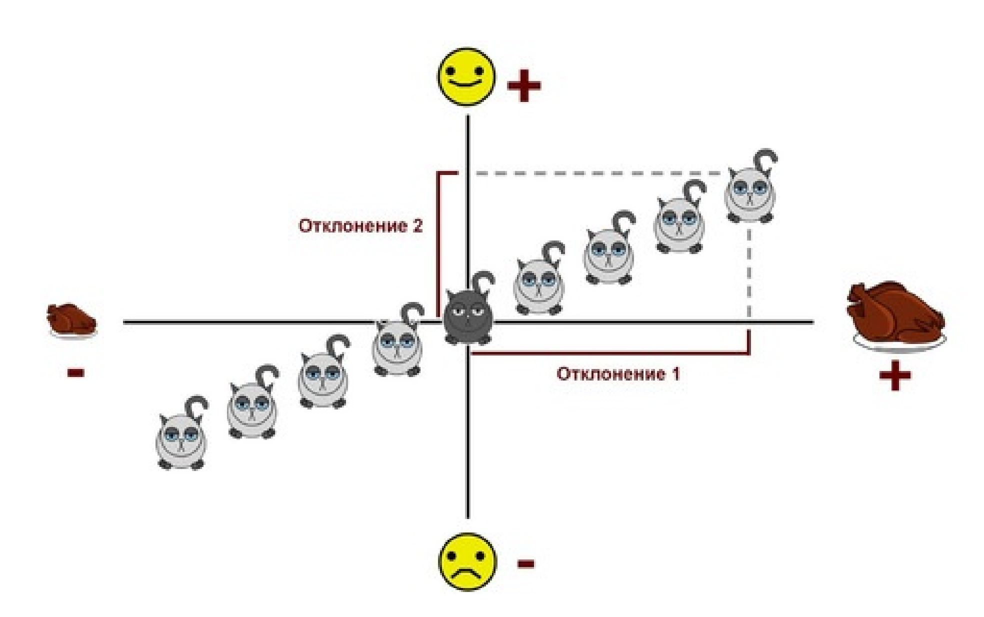


И поэтому мы нуждаемся в некоторой мере, которая позволила бы нам, во-первых, оценить, насколько сильно связаны между собой счастье и количество доступной еды, а во-вторых, является ли эта связь положительной или отрицательной.

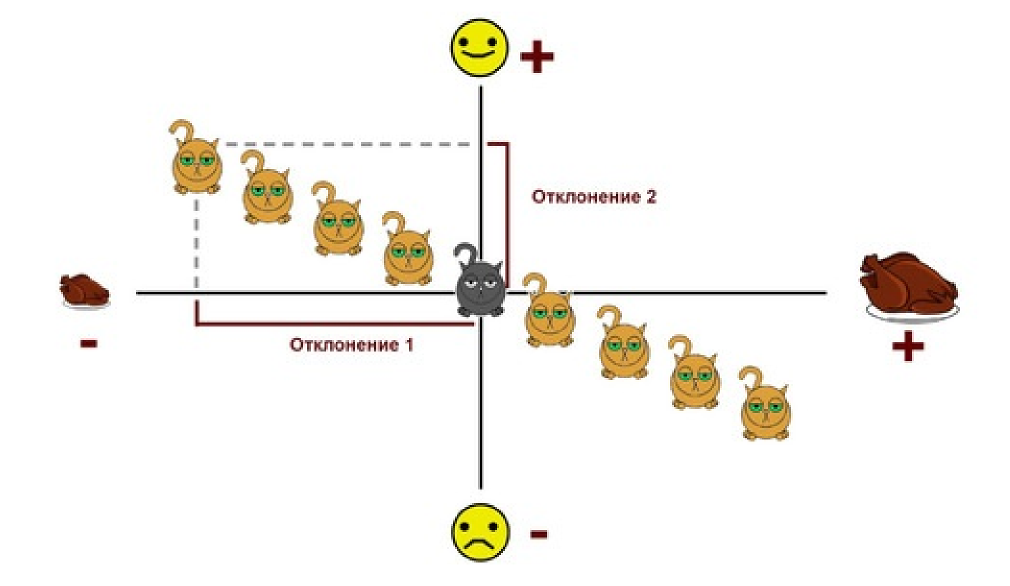
Для вычисления такой меры воспользуемся хитрым способом. Для начала представим, что у нас наблюдается линейная положительная связь. Теперь посчитаем средние арифметические по размеру порций и уровню счастья, а затем возьмем эти показатели в качестве нулевых точек отсчета для нашего графика. После этого мы можем увидеть, что часть котиков более счастлива и получает больше еды, чем в среднем, а остальные — менее счастливы и получают меньше еды, чем средний котик.



Отклонения от среднего по обеим величинам у первых, зажиточных котиков будут положительными числами, а у вторых — отрицательными. Однако если вы возьмете любого из них (назовем его Барсиком) и перемножите его отклонения между собой, то вы получите положительное число. В том числе и потому, что минус на минус дает плюс.

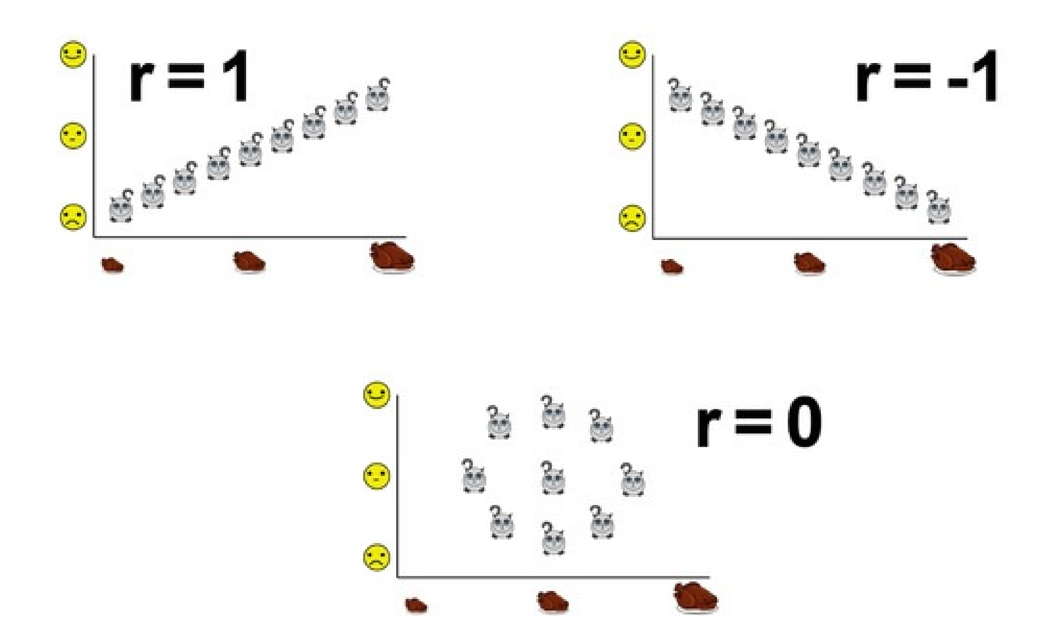


Теперь представим обратную ситуацию: чем больше порции, тем менее счастливыми становятся котики (типичного представителя этой группы мы назовем Мурзиком). В этом случае мы также наблюдаем разделение на две группы: несчастных обжор и счастливых голодающих. Но и у тех, и у других знак одного отклонения будет положительным, а знак другого — отрицательным. А как мы знаем, произведение положительного и отрицательного чисел дает отрицательное число.



Иными словами, знак, который получается при перемножении отклонений, может служить индикатором того, является ли наш котик Барсиком, который становится счастливее при увеличении порций, либо Мурзиком, которому еда отвратительна. Осталось только понять, кто из них делает больший вклад в наблюдаемые данные, что достигается простым суммированием полученных произведений. Если при результате стоит плюс, то победили Барсики и связь положительная. Если минус — то преобладают Мурзики и связь отрицательная. Если же ответ близок к нулю, объявляется боевая ничья и признается отсутствие связи.

Далее с помощью некоторых нехитрых преобразований этот результат приводят в нужную размерность, получив так называемый *коэффициент корреляции Пирсона*. Он может изменяться в пределах от -1 до 1, где -1 — отрицательная связь, +1 — положительная связь, а 0 — отсутствие всякой связи.



Нулевая гипотеза такого коэффициента — связи нет, альтернативная — связь есть (не важно, положительная или отрицательная). Если коэффициент корреляции достаточно большой по модулю, то нулевая гипотеза отвергается в пользу альтернативной.

Основная проблема r Пирсона как параметрического критерия (т. е. использующего в расчетной формуле средние значения) заключается в том, что он очень не любит выбросы и ненормальные распределения. Поэтому у него есть непараметрический аналог —

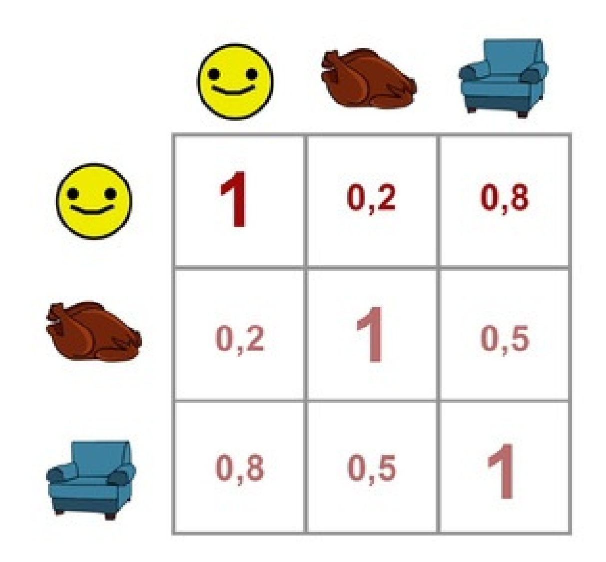
коэффициент корреляции Спирмена.

Чтобы его вычислить, упорядочим наших котиков от самого счастливого до самого несчастного и присвоим им ранги. Затем мы перераспределим их от самого переедающего до самого голодного и присвоим им ранги уже по этому признаку. Если результаты обоих ранжирований будут совпадать между собой, то мы можем констатировать положительную связь, если же они будут диаметрально противоположными — отрицательную.



Критерий Спирмена мы получаем, применив специальную формулу к нашим рангам, и он интерпретируется аналогично r-критерию Пирсона.

Как правило, проводя корреляционный анализ, мы анализируем сразу несколько переменных и по итогу получаем так называемую корреляционную матрицу. В ней записаны все вычисленные коэффициенты корреляции. Чтобы найти, какие переменные связаны с счастьем, достаточно найти нужный столбик и посмотреть, какие из этих коэффициентов являются значимыми.



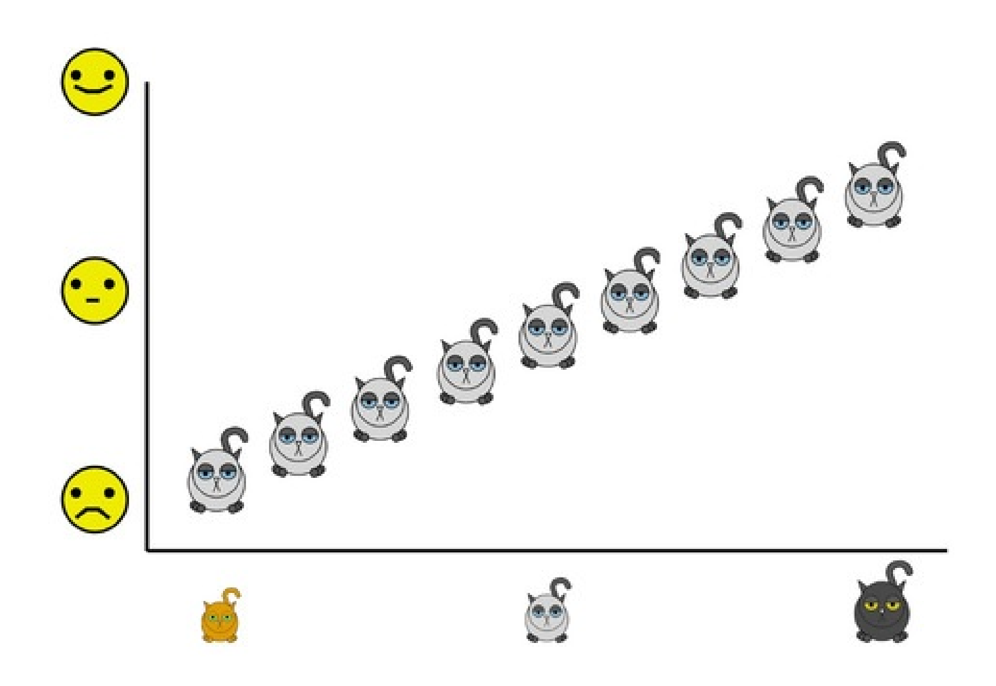
Единственное — если вы находите несколько коэффициентов корреляции одновременно, то

здесь опять возникает проблема множественных сравнений. Решить ее можно, применив всю ту же поправку Бонферрони: поделив критический p-уровень значимости (0,05) на количество вычисленных критериев (в нашем случае на 3) и сравнив наш p-уровень с получившимся значением (0,017).

К большому сожалению, корреляционный анализ позволяет установить только само наличие связи. Однако сказать, насколько сильно тот или иной фактор влияет на счастье, он не способен. Для этого используются более мощные методы, о которых мы поговорим в следующей главе.

НЕМАЛОВАЖНО ЗНАТЬ!
Корреляция может обмануть

При проведении корреляционного анализа очень важно помнить, что высокий коэффициент корреляции не всегда указывает на характер связи между явлениями. В качестве примера предположим, что мы нашли взаимосвязь между размером котиков и их эмоциональным состоянием. Иными словами — чем больше котик, тем он счастливее.



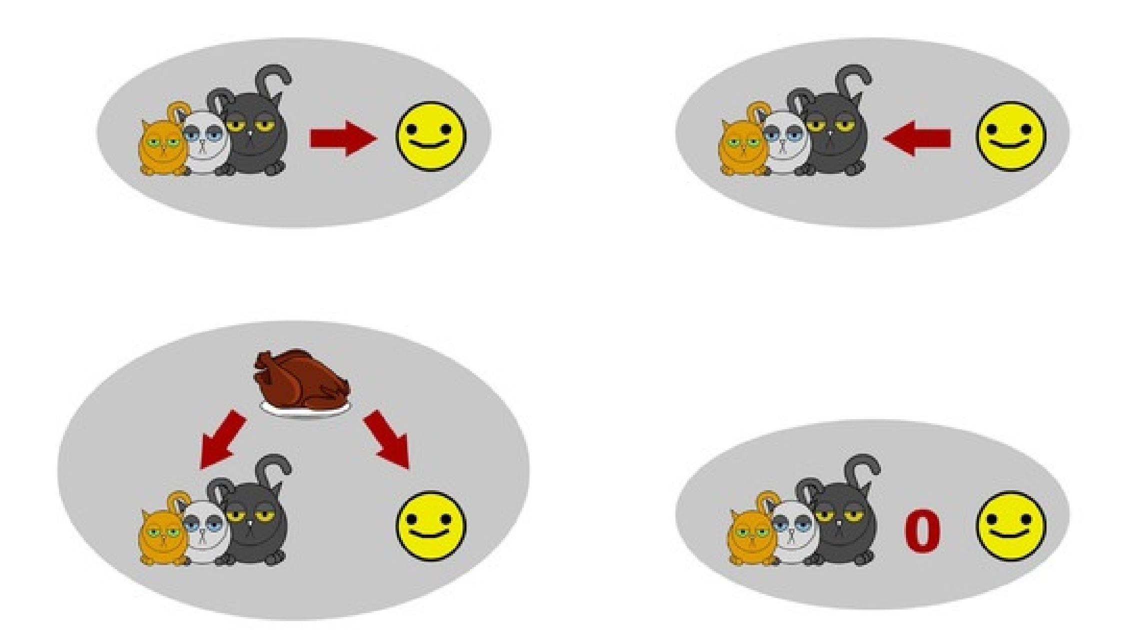
Тогда теоретически равноправными являются следующие утверждения.

1. Большие котики лучше реализуются в жизни и от того более счастливы.

2. Хорошее расположение духа вызывает более активную выработку гормонов роста, что и приводит к данному эффекту.

3. Существует некоторая третья переменная, которая обусловливает как хорошее настроение, так и разницу в размерах. Например, качество и количество котикового корма.

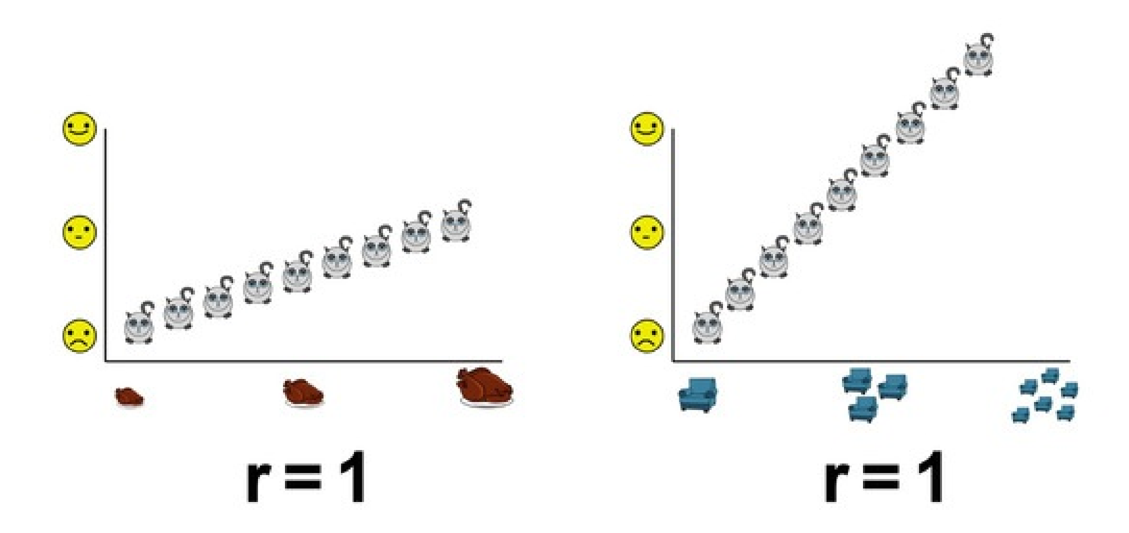
4. Это просто совпадение.



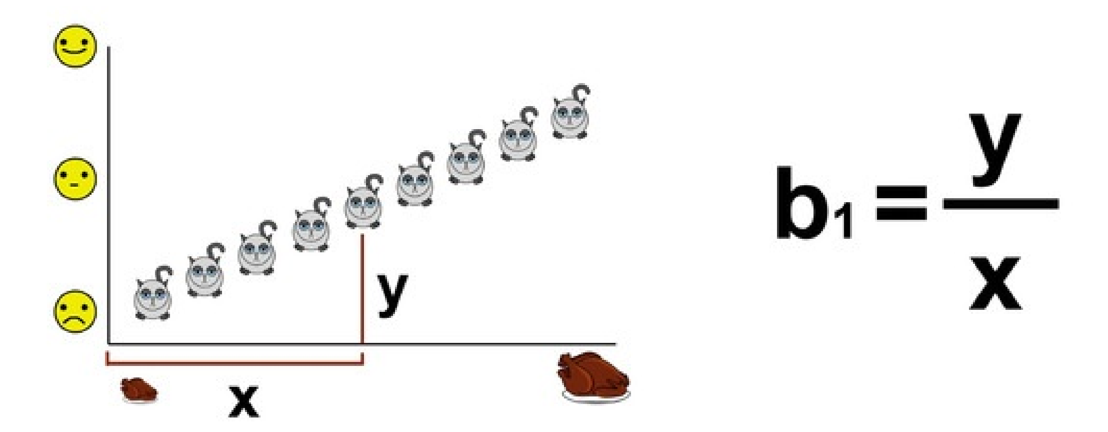
И чтобы определить, какая из этих гипотез верна, необходимо организовать экспериментальное исследование, о котором шла речь в предыдущих главах.

Глава 10. Формула счастья или основы регрессионного анализа

Из предыдущей главы вы узнали, как определить, что делает наших котиков счастливыми. Для этих целей мы использовали корреляционный анализ. Однако коэффициенты корреляции позволяют установить лишь само наличие и выяснить направление этой связи. Определить, насколько сильно изменяется одна переменная под воздействием другой, он не в силах. В качестве иллюстрации приведем пример.

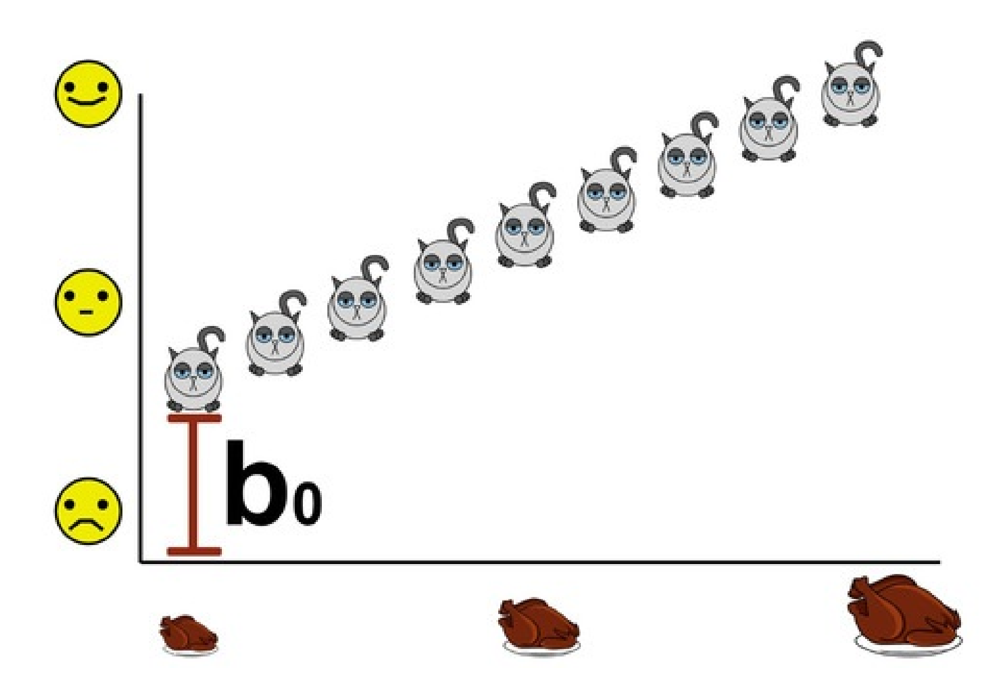


На графиках изображены две линейные положительные взаимосвязи. Коэффициент корреляции в обоих случаях равен +1. Однако очевидно, что каждый подранный диван делает котиков гораздо счастливее, чем очередное увеличение пайков. Эта разница математически описывается с помощью коэффициента b_1. Он определяется как тангенс угла между линией котиков и горизонтальной оси x. Чем больше этот коэффициент, тем сильнее растет уровень счастья от каждой новой порции.



Можно выразиться и так: при увеличении порции мяса на одну единицу котиковое счастье будет возрастать на b_1.

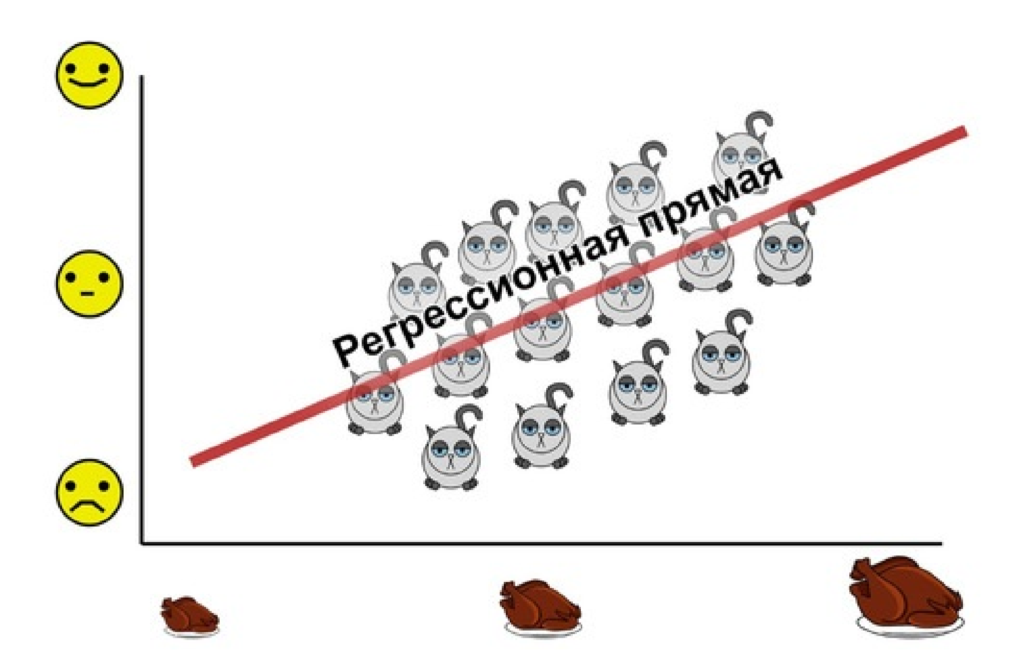
Вторая величина, которая может описывать нашу прямую, называется b_0. Она показывает, насколько счастливы котики, если их совсем не кормить.



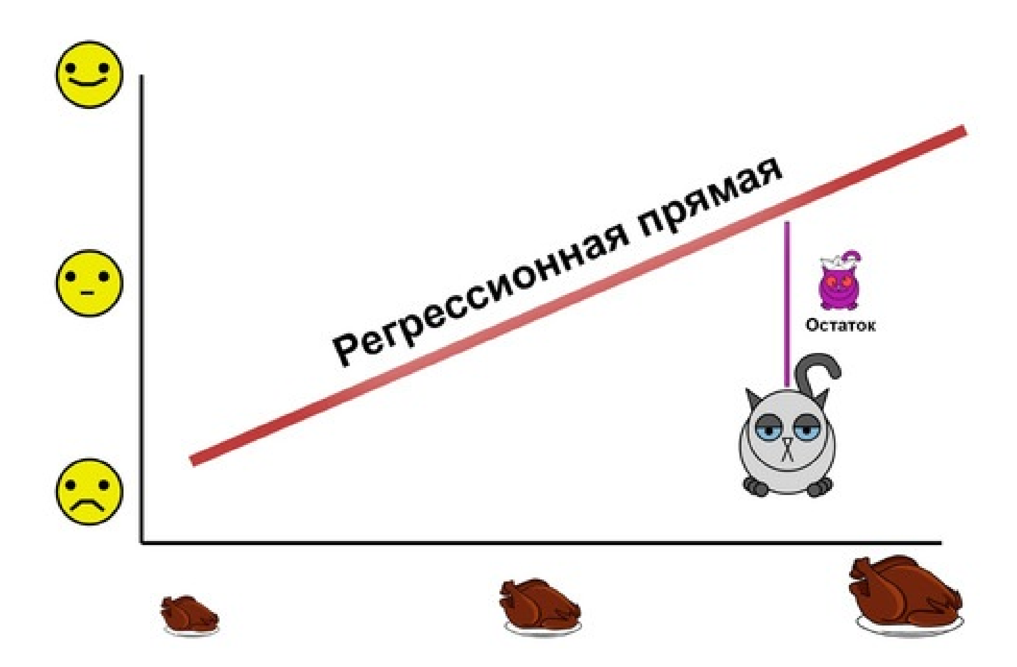
По итогу, линейную взаимосвязь между количеством еды и котиковым счастьем можно описать с помощью вот такого несложного уравнения.

$$= b_0 + b_1 \times$$

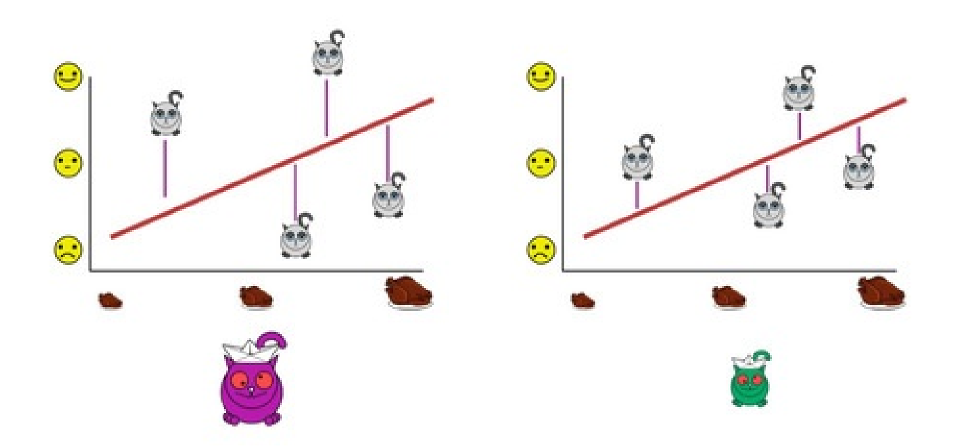
Однако, к сожалению, реальные взаимосвязи мало похожи на прямую линию. Чаще они напоминают собой огурец, а в запущенных случаях — авокадо. Но описывать такие вещи довольно сложно, поэтому статистиками был разработан специальный метод, который позволяет подобрать такую прямую, которая смогла бы заменить этот овощ с минимальными потерями данных. Этот метод называется *регрессионным анализом*, и результатом его применения обычно является уравнение, похожее на то, что обозначено нами выше.



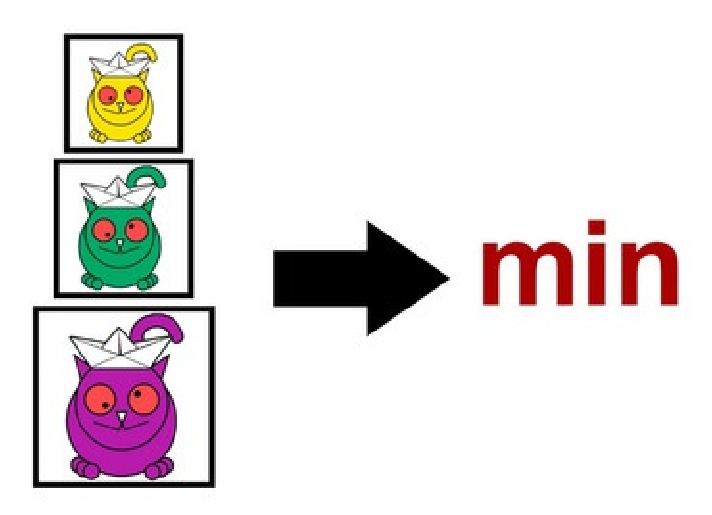
Рассмотрим, как это получается. Предположим, у нас есть прямая, полученная в результате регрессионного анализа, и недалеко от этой прямой обосновался наш старый знакомый — Барсик. На рисунке видно, что Барсик чуть менее счастлив, чем ему положено при своем рационе. Это различие называется *регрессионным остатком*.



Теперь мысленно подвигаем Барсика относительно регрессионной прямой — при удалении от нее остаток будет увеличиваться, а при приближении — уменьшаться. И, наконец, если Барсик встанет на эту прямую, остаток будет равен нулю. А теперь вспомним, что у нашего Барсика есть компания, и если все наши котики находятся на прямой, то их совокупный остаток тоже будет равен нулю. В то же время при удалении от этой прямой совокупный остаток начнет увеличиваться.



Логика диктует, что, чтобы получить такой совокупный остаток, нам нужно просто сложить индивидуальные остатки котиков (бр-р-р… звучит жутко). Однако, поскольку эти остатки могут быть как положительными, так и отрицательными (некоторые котики ведь могут быть более счастливыми, правда?), на выходе мы можем получить полную белиберду (аналогичная ситуация была, когда мы считали стандартное отклонение). Поэтому, чтобы исключить влияние знаков, мы складываем квадраты остатков.



Чем больше получившаяся сумма, тем хуже прямая описывает наши данные. И суть регрессионного анализа заключается в том, чтобы подобрать такую прямую, при которой эта сумма была бы минимальной.

А теперь пару слов о том, почему регрессионный анализ считается одним из самых крутых статистических методов. Дело в том, что он способен работать с большим количеством переменных одновременно. И если вы умудритесь провести тотальный замер ваших котиков на предмет того, что может приносить им счастье, и прогоните эти данные через регрессионный анализ, вы можете получить настоящую формулу счастья.

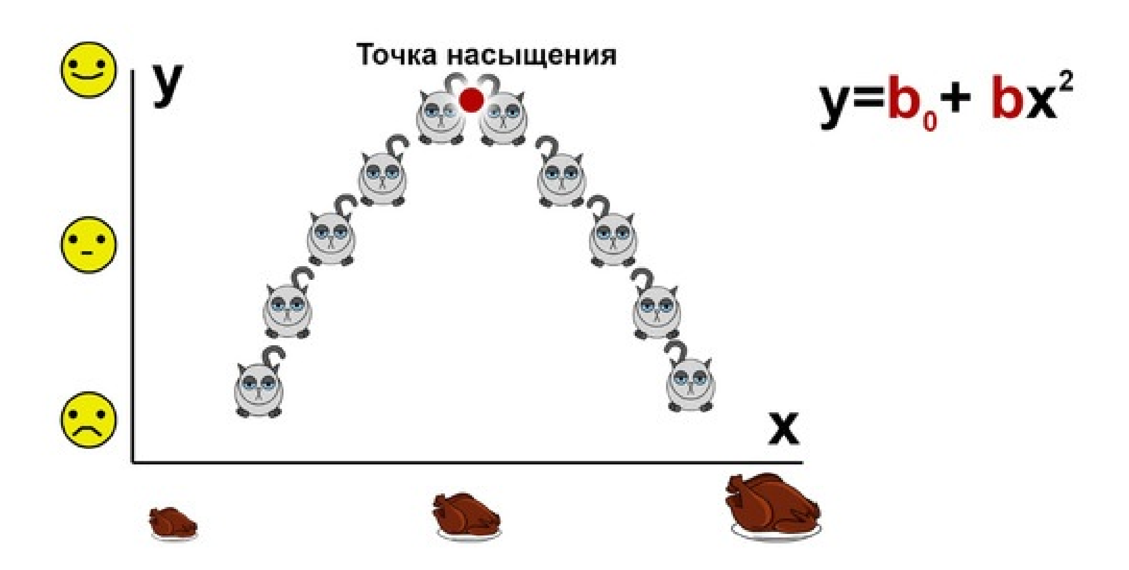
По этой формуле вы сможете выяснить, какие факторы наиболее сильно влияют на котиковое счастье, и предсказывать, насколько будет счастлив тот или иной котик по их значениям.

Однако здесь важно сделать предостережение — если вы вычислили такую формулу, это вовсе не означает, что то, что в ней справа — причины, а слева — следствие. В конце концов, может быть, еда делает котиков счастливыми, а может, и наоборот — у счастливых котиков лучше аппетит.

Помимо самой формулы вы также можете получить информацию о том, можно ли в нее что-нибудь добавить. В этом вам поможет коэффициент детерминации R^2. Он изменяется в промежутках от 0 до 1, и чем ближе к единице, тем лучше ваша формула объясняет наблюдаемые данные. Низкий коэффициент детерминации говорит о том, что нужно поискать, какие еще переменные могут быть связаны с котиковым счастьем.

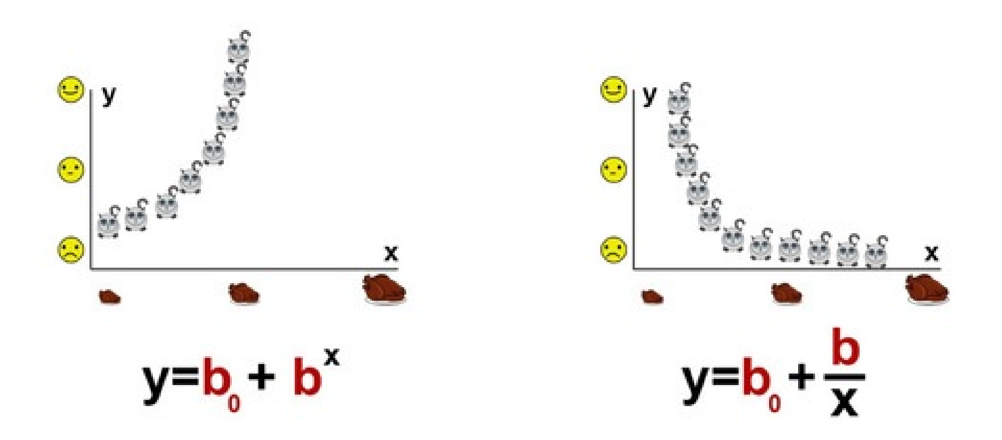
НЕМАЛОВАЖНО ЗНАТЬ!
Нелинейная регрессия

Вообще-то говоря, связь между переменными не всегда является линейной. Например, существует определенный момент, после которого котика начинает тошнить от дополнительных порций, хотя до этого момента каждая новая порция делала его более счастливым.



Такую взаимосвязь можно описать с помощью *квадратного* (или, как говорят математики, *полиномиального*) *уравнения*, с которым мы знакомы со школы. И составить такое уравнение можно с помощью метода *полиномиальной регрессии*.

Определить целесообразность использования этого или сходных с ним методов можно, предварительно построив точечные диаграммы. Помимо линейных и полиномиальных взаимосвязей могут быть еще и такие.

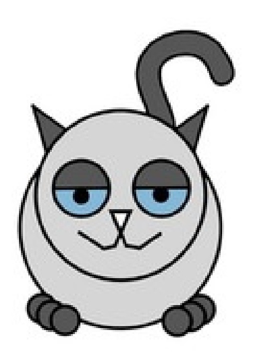


Увидев, что ваша взаимосвязь похожа на что-нибудь из этого, вы можете либо найти


подходящий метод регрессионного анализа, либо преобразовать одну из переменных таким образом, чтобы можно было бы воспользоваться методами линейной регрессии.

Глава 11.
Котики счастливые и несчастные или логистическая регрессия и дискриминантный анализ

Из предыдущей главы вы узнали, как с помощью линейной регрессии понять, насколько сильно те или иные факторы влияют на уровень котикового счастья. Однако, у обычного регрессионного анализа есть одно существенное ограничение — уровень счастья должен быть достаточно точно измерен с помощью какого-нибудь прибора или теста. К сожалению, мы зачастую не располагаем подобным оборудованием. Максимум, что мы можем сделать, это прикинуть, является ли данный конкретный котик счастливым или несчастным.



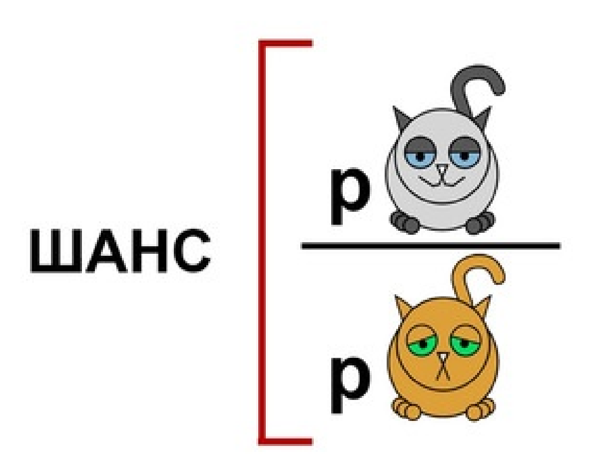


Можем ли мы при таких условиях найти факторы, предсказывающие котиковое счастье?

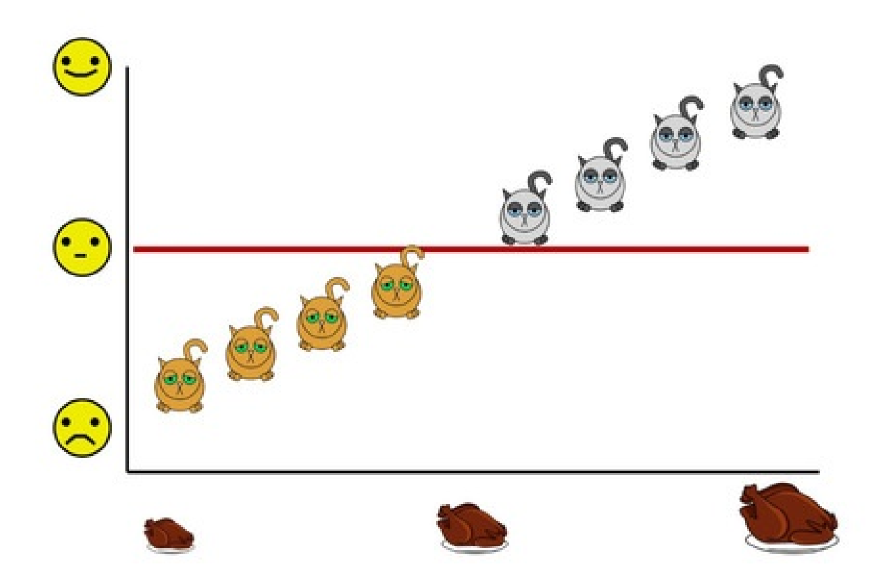
Разумеется да. И для этого существуют два очень хороших метода. Первый называется *логистической регрессией*, а второй — *дискриминантным анализом*.

Логистическая регрессия во многом похожа на линейную. Однако вместо уровня счастья в левой части уравнения стоит величина, которая позволяет рассчитать вероятность того, что данный котик счастлив. Эта величина называется *логарифмом шанса*.

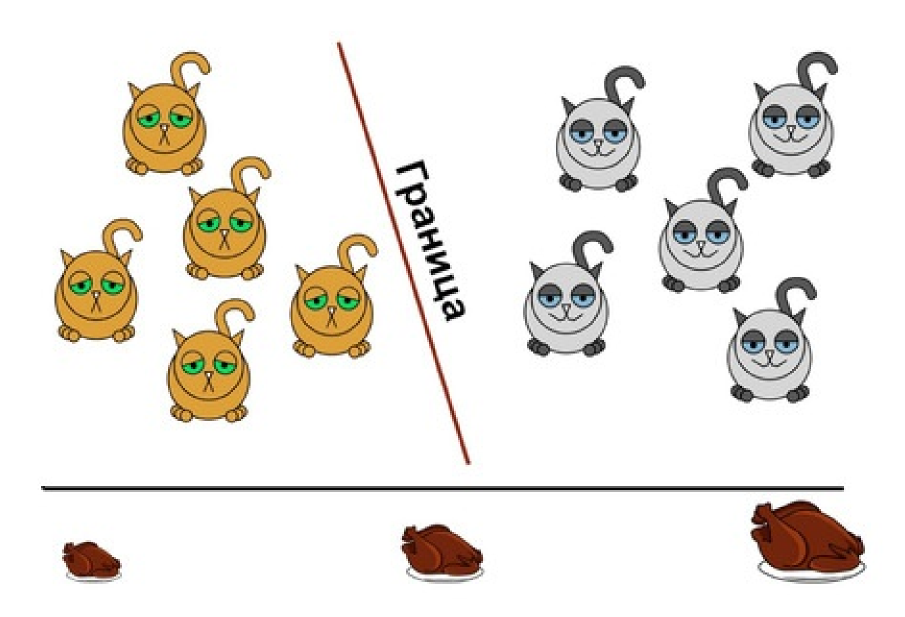
Слово «шанс» достаточно часто встречается в русском языке, как правило, обозначая то, что ни в коем случае нельзя упустить. Но с точки зрения статистики шанс — это вероятность того, что данный котик счастлив, деленная на вероятность того, что он несчастлив.



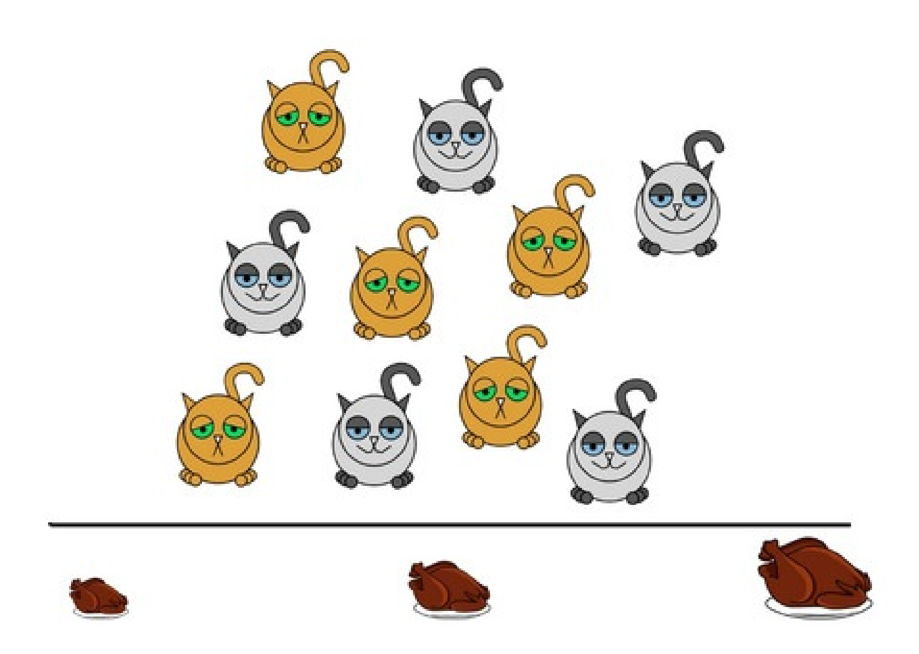
По некоторым математическим причинам от шанса берут натуральный логарифм и подставляют эту величину в регрессионное уравнение. Если логарифм шанса будет положительным, то данный котик считается счастливым, а если отрицательным — то несчастным.



Альтернативным методом является дискриминантный анализ. Чтобы разобраться, что это такое, обратимся к рисунку.



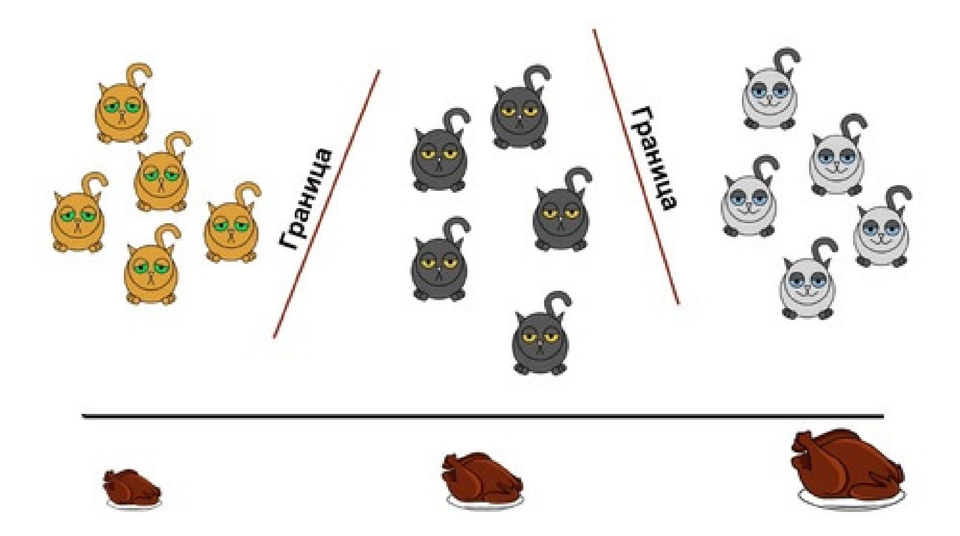
На нем представлены счастливые котики (Барсики) и несчастные (Мурзики), а также информация о том, кто из них сколько ест. Очевидно, что Барсики едят в целом больше, и мы можем провести четкую границу между котиками по этому фактору. И если такая граница возможна, то мы делаем вывод, что фактор связан с уровнем счастья. Иной случай выглядит так.



Здесь невозможно построить такую границу, чтобы Барсики оказались по одну ее сторону, а Мурзики — по другую. Соответственно, в этом случае количество еды не связано с уровнем счастья.

Алгоритм нахождения таких границ и называется дискриминантным анализом, а формула, которая задает границы, — *дискриминантной функцией*. По итогу дискриминантного анализа вы получаете таблицу, в которой обозначается, по каким факторам удалось провести внятные границы, а по каким — нет.

Дискриминантный анализ может работать и с большим количеством групп. Например, если мы добавим к нашим Барсикам и Мурзикам группу философских котиков, дискриминантный анализ сможет найти границы между ними всеми. Число таких границ всегда будет на одну меньше, чем количество групп.



Если же вы являетесь поклонником регрессионного анализа, то при большом количестве групп вы можете вычислить так называемую мультиномиальную регрессию.

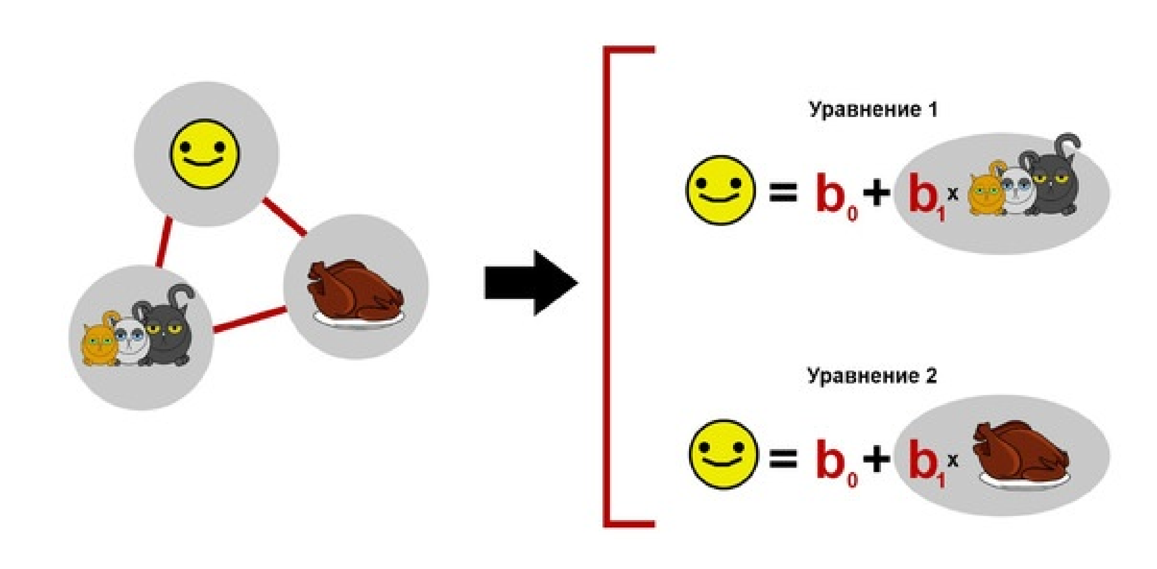
НЕМАЛОВАЖНО ЗНАТЬ!
Мультиколлиниарность и переобучение

С методами регрессионного и дискриминантного анализов связаны две проблемы, которые существенным образом могут испортить вам все ваши выводы.

Первая из них — *проблема мультиколлиниарности* — возникает в случаях, когда некоторые факторы сильно коррелируют между собой, и приводит к неустойчивости получившегося уравнения. Проявляется это в двух формах.

1. При добавлении всего одного-двух котиков в выборку это уравнение может измениться до неузнаваемости.

2. Формулы, построенные на двух сходных выборках котиков, будут различаться.



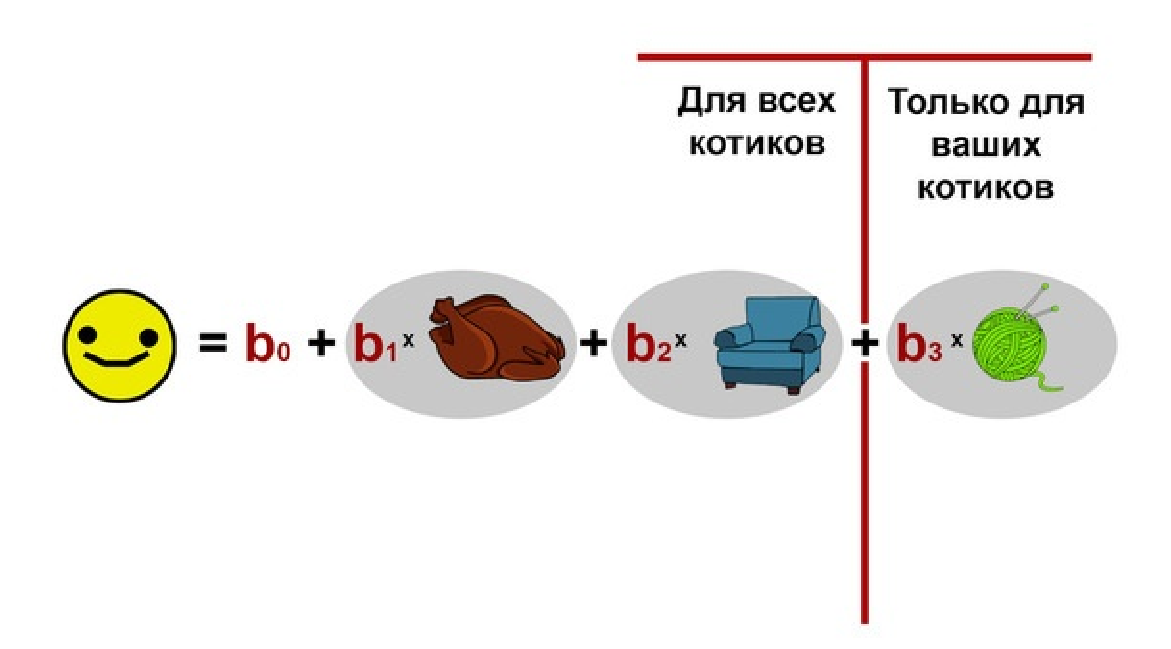
Как правило, эту проблему преодолевают тремя способами.

1. Исключают одну из коррелирующих переменных из анализа.

2. Предварительно проводят процедуру *факторного анализа* (о нем будет рассказано далее), заменяющего эти переменные одной искусственной, которая и будет включена в регрессию.

3. Проводят процедуру *пошаговой регрессии*. Такая регрессия постепенно включает в уравнение по одной переменной и сразу же после этого пересчитывает вклад всех остальных. В итоге если одна из коррелирующих переменных была выбрана в качестве фактора, вторая туда скорее всего не попадет.

Вторая проблема — *проблема переобучения* — заключается в том, что уравнение, полученное на одних котиках, может не работать на других. Она возникает из-за того, что в вашей выборке котиков могут быть закономерности, которые нехарактерны для котиков в целом. И зачастую они попадают в регрессионную модель.

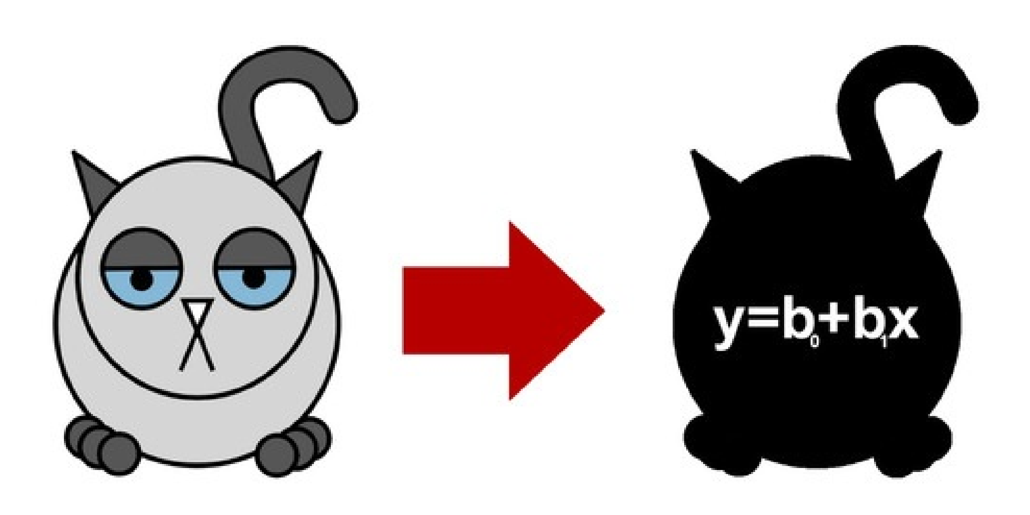


Для того чтобы предотвратить переобучение, используют критерий, который искусственно ограничивает количество факторов, включенных в уравнение (например критерий Акаике и Байесовский информационный критерий).

Глава 12.
Котиковые аналоги
или основы математического моделирования

В предыдущих разделах мы подробно рассмотрели метод регрессионного анализа, который позволяет построить уравнение, описывающее, как различные вещи влияют на настроение котиков. Подобные уравнения входят в группу объектов, называющихся *математическими моделями*.

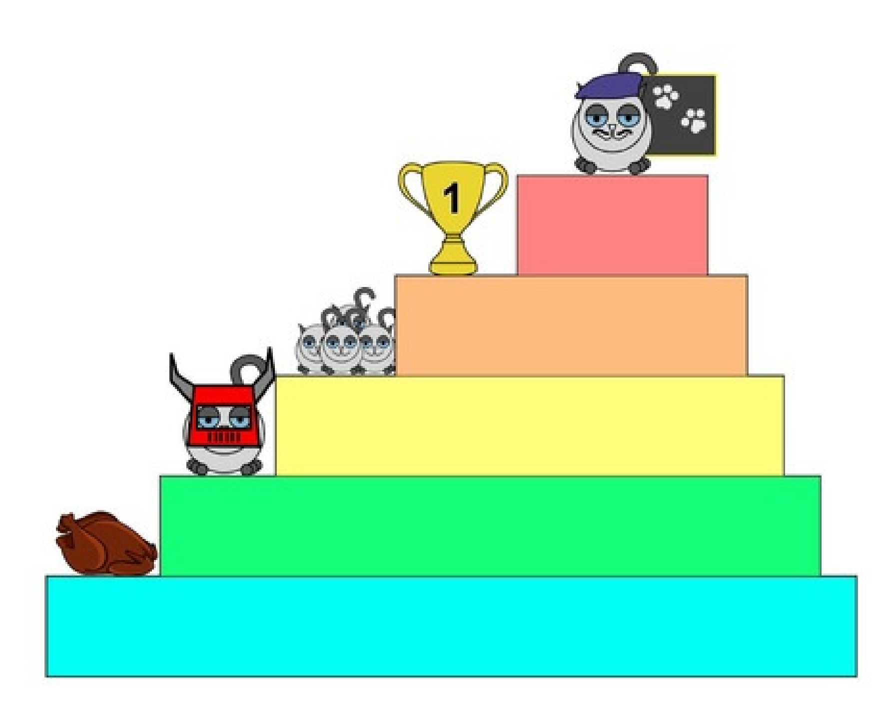
Математическая модель — это своего рода аналог котика, который позволяет изучать его поведение без проведения реальных экспериментов. Как правило, это значительно удешевляет исследования.



Все математические модели делятся на *функциональные* и *структурные*. Функциональные модели, к которым, к слову, относится регрессионное уравнение, — описывают влияние внешних факторов на котиковое состояние. Например, известная нам модель котикового счастья.

$= b_0 + b_1 \times \quad + b_2 \times \quad + b_3 \times$

Особенность такой модели в том, что мы подробно не рассматриваем состав этого счастья. Счастье для нас — некий целостный объект, целевая переменная, которая может меняться: прибывать или убывать. А вот структурные модели позволяют описать его компоненты: от удовлетворения базовых котиковых потребностей до котиковой самореализации.



Как правило, функциональные модели записываются с помощью уравнений. А вот структурные могут быть достаточно разнообразными: от таблиц до блок-схем.

Любая математическая модель строится в два этапа. На первом этапе мы прикидываем, какие факторы в принципе могут влиять на котиковое счастье или из каких компонентов оно может состоять. Этот этап называется также *построением содержательной модели.*

Второй этап включает в себя сбор реальных данных и их математическую обработку. Он называется *построением формальной модели*. Формальную модель уже можно использовать как аналог реального котика. Изменяя различные параметры этой модели, вы сможете понять, как функционирует котик, не прибегая к опытам над животными.

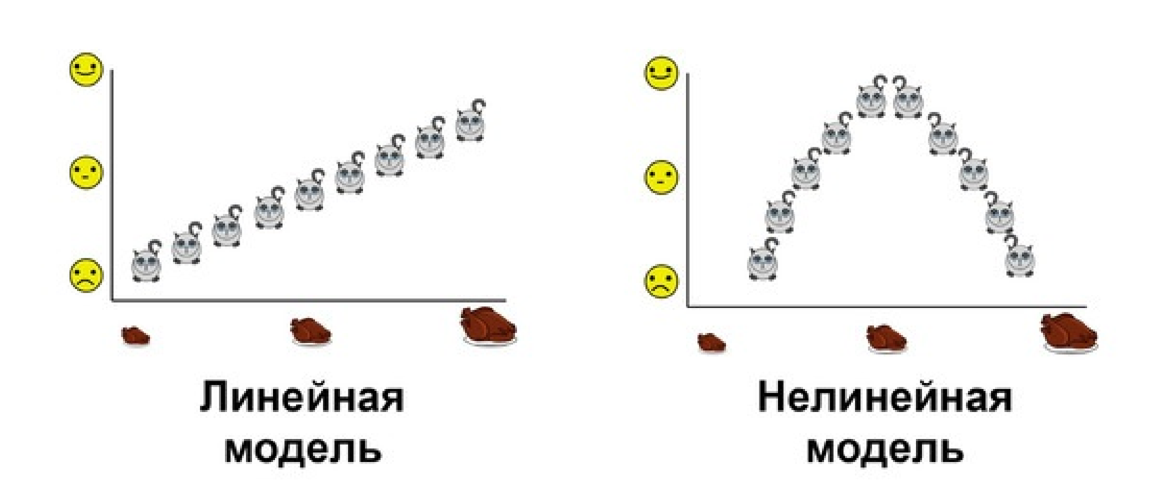
НЕМАЛОВАЖНО ЗНАТЬ!

Классификация математических моделей

Помимо деления на функциональные и структурные модели есть еще несколько классификаций, о которых полезно знать. В частности бывают модели *статические* и *динамические*. Первые описывают состояние котика в какой-то конкретный момент. Вторые же концентрируются непосредственно на изменениях, которые претерпевает котик.



Кроме того, модели делятся на *линейные* и *нелинейные*. Линейные модели включают в себя только линейные взаимосвязи, о которых мы подробно говорили в главах про корреляционный и регрессионный анализы. Нелинейные модели могут включать в себя нелинейные взаимосвязи. Примером здесь может служить полиномиальная регрессия.



Также имеет смысл рассмотреть деление моделей на *непрерывные* и *дискретные*. Первые отличаются тем, что в них все переменные имеют бесконечное множество значений. Пример такой переменной — это котиковый размер, измеренный в сантиметрах. Мы можем сказать, что наш котик имеет длину 62 см. А можем — что 62,513987 см. И даже точнее. Если состояние

вашего котика измеряется такой переменной, то, чтобы построить функциональную модель, вам необходима линейная регрессия.

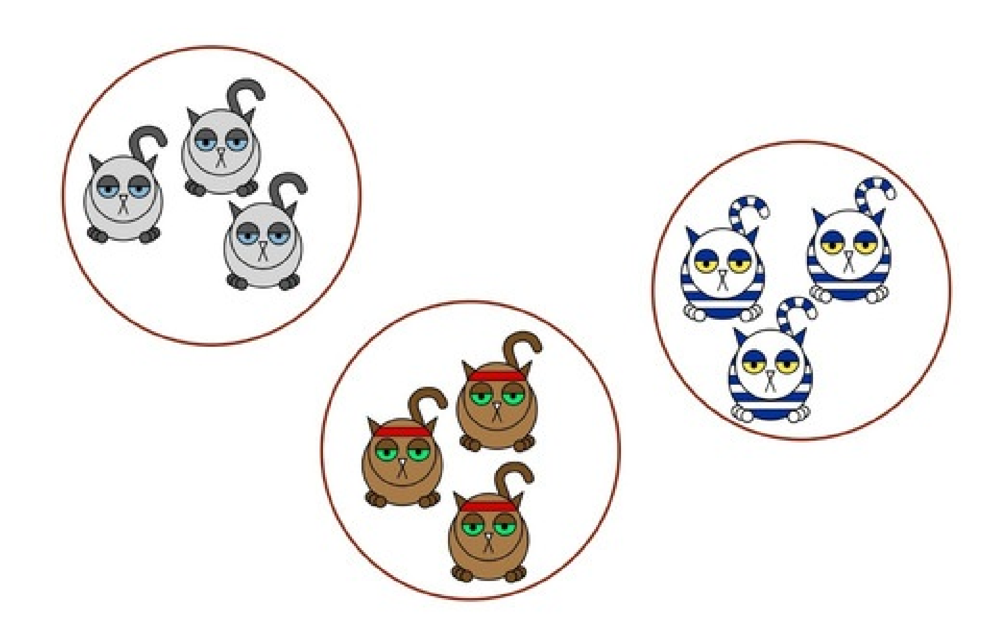
Дискретные же модели работают с переменными, которые имеют ограниченное количество значений. Например, тот же размер, но имеющий только три значения: маленький, средний и большой. Построить модели с дискретными целевыми переменными, в частности, позволяют логистическая регрессия и дискриминантный анализ.



Впрочем, на практике большинство моделей относятся к смешанным типам — в них встречаются как дискретные, так и непрерывные переменные, а линейные взаимосвязи вполне могут сочетаться с нелинейными.

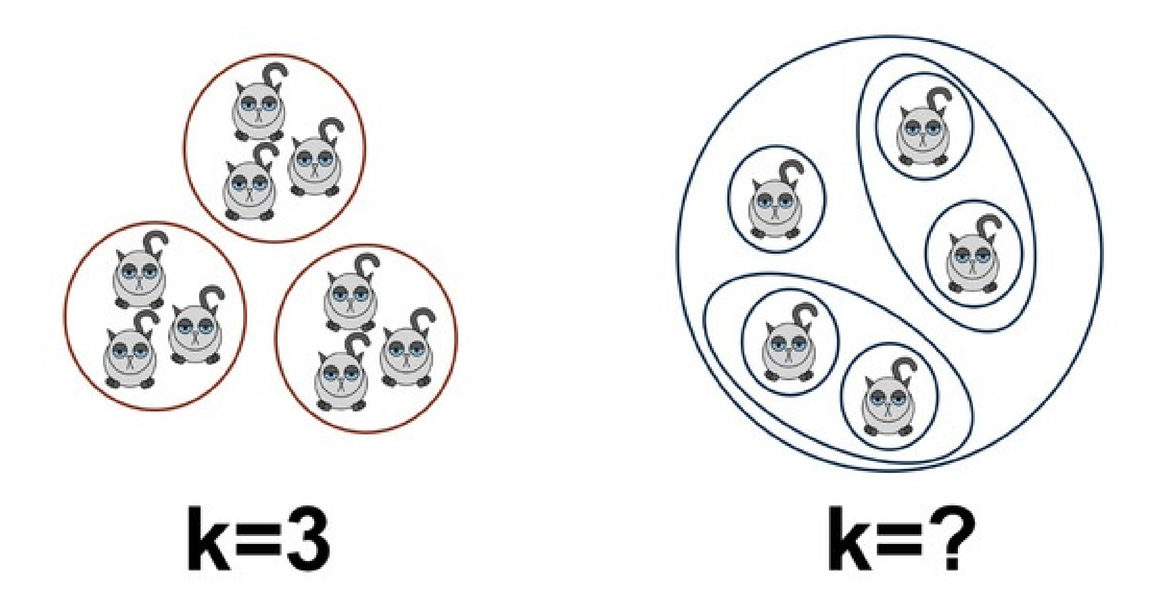
Глава 13. Разновидности котиков или основы кластерного анализа

Из предыдущих разделов мы узнали, как определить, какие факторы делают наших котиков счастливыми. В этом нам помогли регрессионный и дискриминантный анализы. Зная значения этих факторов, мы можем предсказать, будет ли тот или иной котик счастливым или несчастным. Иными словами, мы можем рассортировать котиков по классам, т. е. *классифицировать* их.



Вообще, задача классификации является крайне важной практически для всех наук, изучающих котиков. Но довольно часто мы не имеем никакого понятия даже о том, на какие группы делятся котики. Ведь котики очень разные. Поэтому существуют методы, которые позволяют не только рассортировывать котиков на группы, но и выделять сами эти группы. И все вместе они называются *кластерным анализом*.

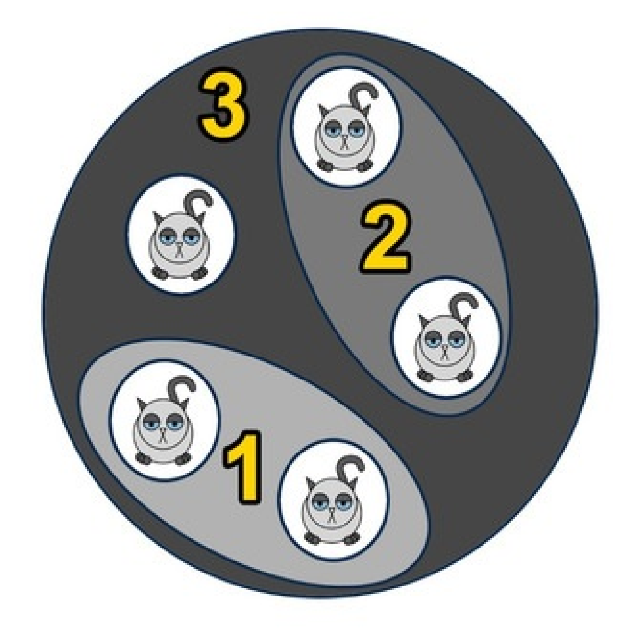
В первом приближении у нас могут возникнуть две ситуации. Первая — мы знаем, на сколько групп у нас должны делиться котики, но не имеем понятия, где эти группы находятся. Вторая — мы не знаем итоговое количество групп. Со второго случая мы, пожалуй, и начнем.



Рассмотрим самый простой пример. Предположим, что мы захотели поделить наших котиков по размеру. Очевидно, что чем больше два котика похожи друг на друга, тем больше шансов, что они окажутся в одной группе. Чтобы понять степень похожести, надо просто найти разность между размерами — чем она меньше, тем более похожими являются наши котики.



Итак, мы вычисляем все возможные разности между размерами котиков. Далее пара самых похожих котиков объединяется в группу (или кластер). Затем мы вновь вычисляем разности. А затем опять объединяем самых похожих. И так происходит до тех пор, пока у нас все котики не объединятся в один большой кластер.



Этот алгоритм относится к методам *иерархической кластеризации*. Их довольно много, но каждый из них обладает следующими свойствами.

1. Эти методы могут работать с большим количеством переменных — вы можете брать и размер, и степень пушистости, и длину коготков, и прочие котиковые признаки одновременно.

2. На основе этих признаков вы вычисляете степень похожести котиков (чаще используется термин *расстояние*).

3. Котики последовательно объединяются в группы. Это может происходить так, как было описано выше (так называемый «*метод ближайшего соседа*»), а может и по другим принципам.

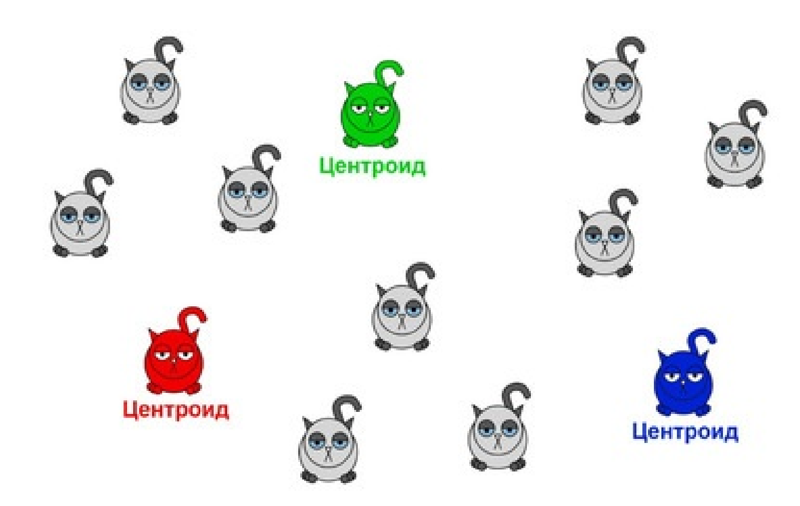
4. По итогу вы получаете график, называемый *дендрограммой*. По ней вы можете определить, на какие группы делятся ваши котики и какие котики к какой группе принадлежат. Единственное — если котиков очень много, воспринимать такую дендрограмму довольно сложно.



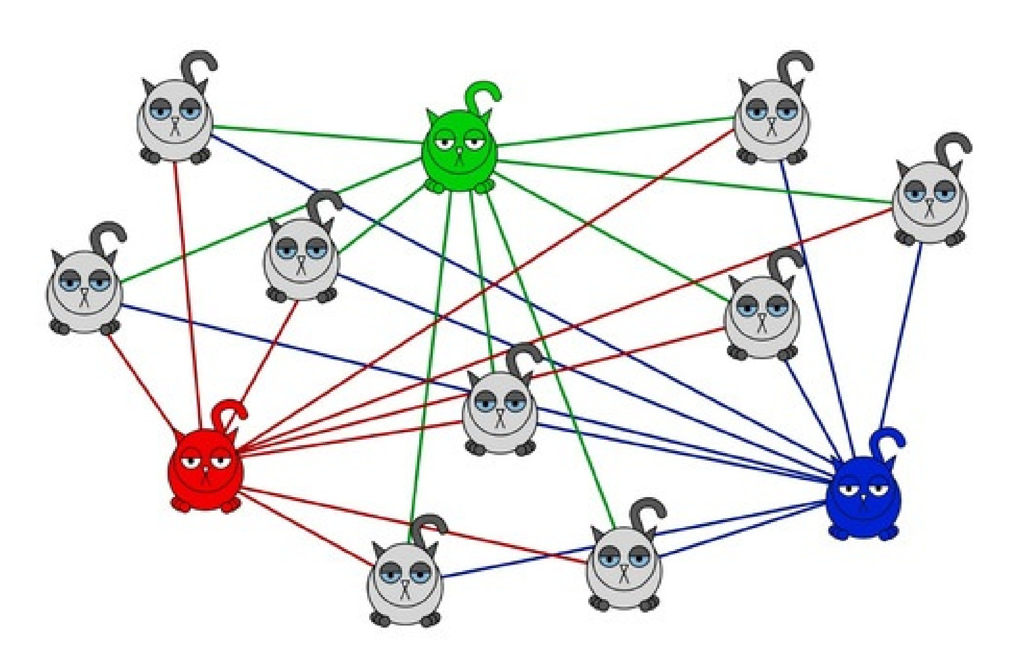
Напомним, что иерархический кластерный анализ позволяет вам разбить котиков на группы, когда вы не знаете, сколько у вас их должно получиться. А если знаете, то более адекватным будет использование метода *k-средних*.

Идея достаточно проста. Предположим, вы подозреваете, что все котики делятся на три различающиеся размером группы. Тогда у каждой группы существует свой представитель, который обладает самым типичным для группы размером. Такой котик называется *центроидом*. И основная задача алгоритма k-средних — найти, каким именно размером эти центроиды обладают.

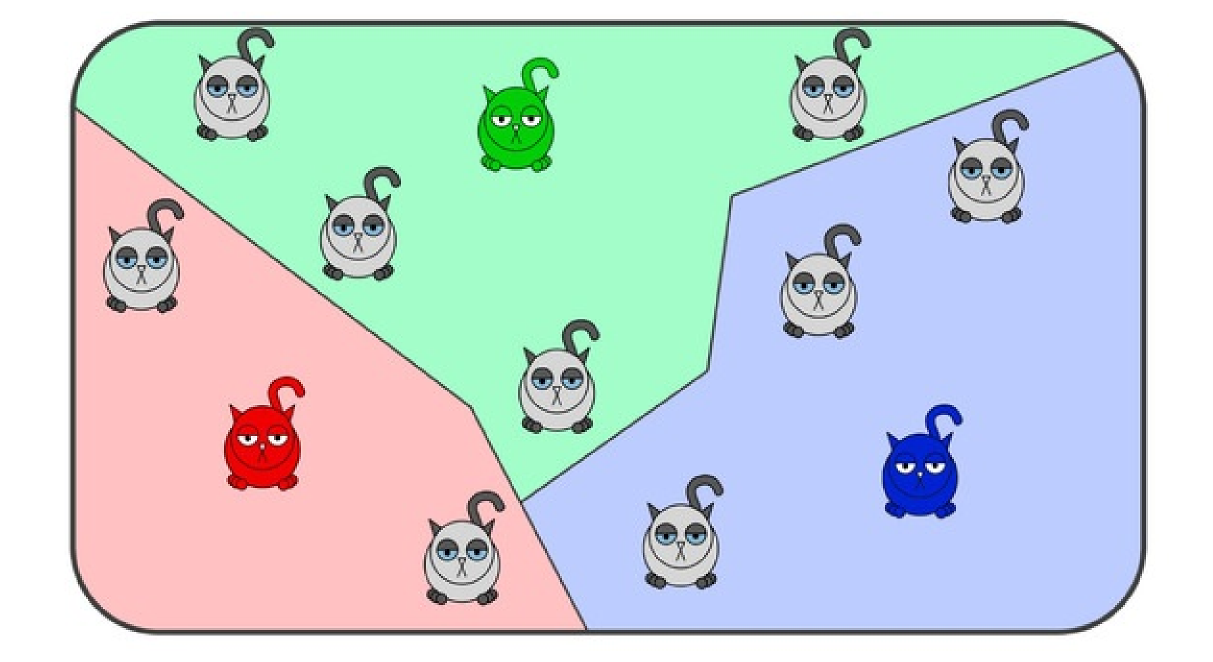
Происходит это пошагово. На первом этапе мы произвольно расставляем центроиды.



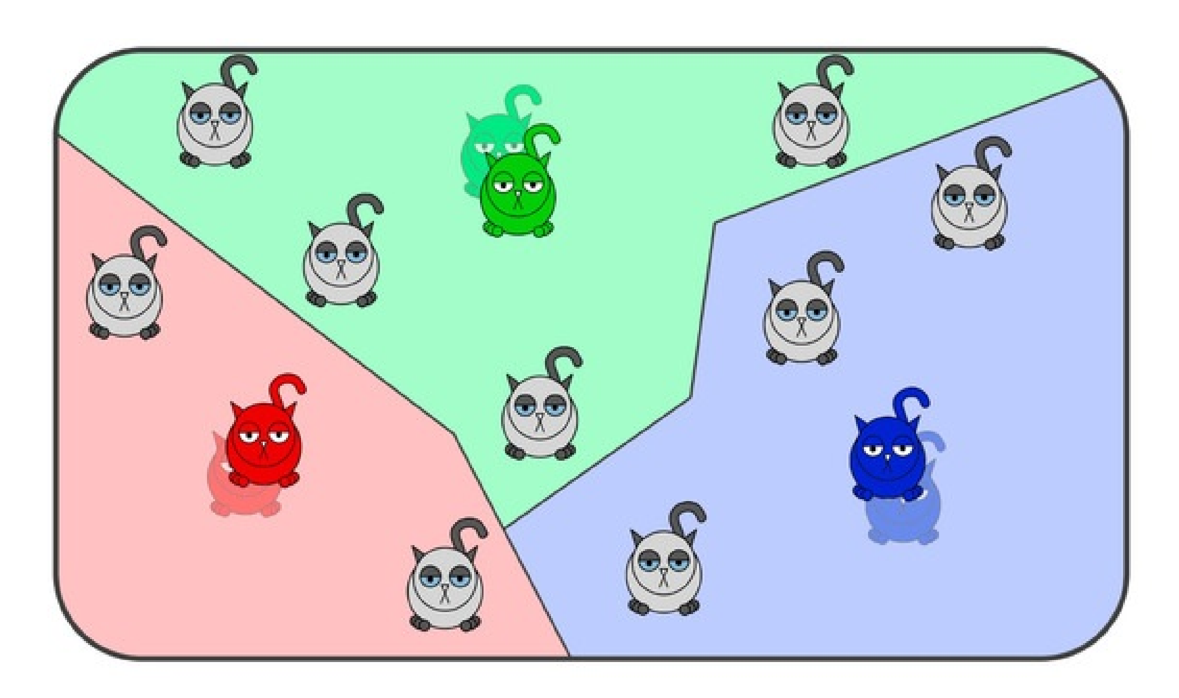
На втором этапе вычисляются расстояния от каждого котика до каждого центроида.



На третьем — определяем принадлежность котиков к тому или иному центроиду. Иными словами — смотрим, какой котик к какому центроиду ближе.

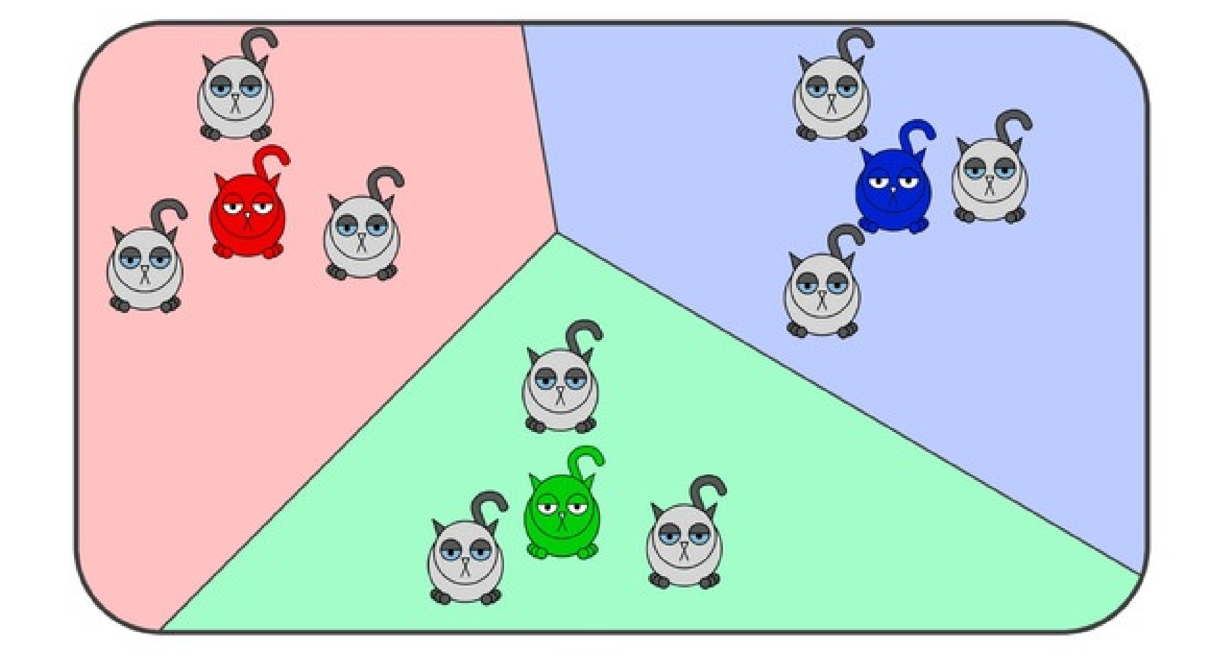


И на четвертом этапе мы вычисляем средний размер котиков при каждом центроиде. И центроид перемещается в этот средний размер.



А потом алгоритм повторяется со второго шага. Происходит это потому, что некоторые котики перебегают от одного центроида к другому, вследствие чего положение центроидов также будет меняться.

Происходит это ровно до тех пор, пока после очередного повторения положение центроидов останется неизменным.

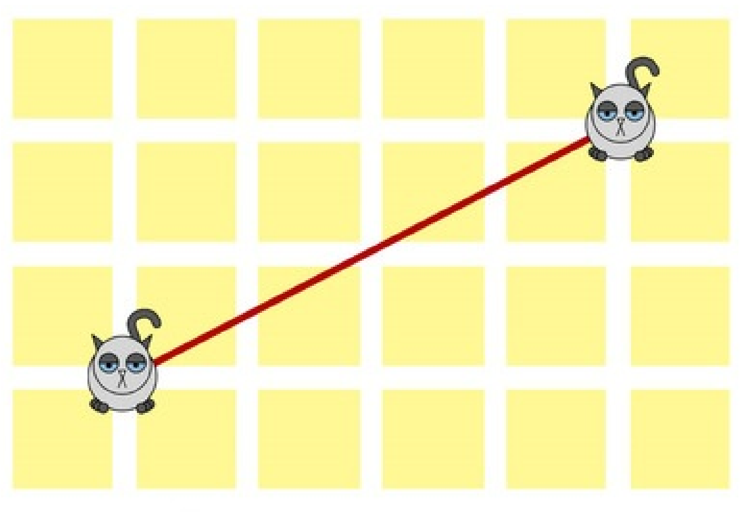


Важно отметить следующие вещи. Во-первых, k-средних может работать сразу по нескольким переменным. Для этого, как и для иерархического кластерного анализа, вычисляется расстояние, но уже не между отдельными котиками, а между конкретным котиком и центроидом.

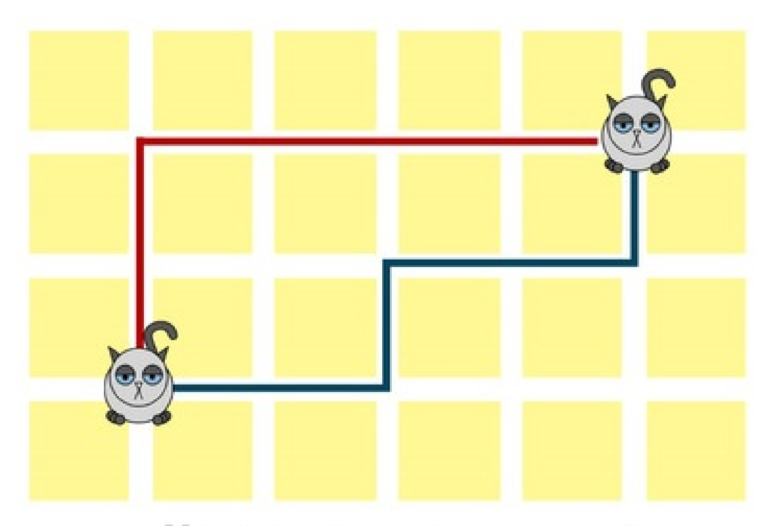
Во-вторых, результат k-средних сильно зависит от начального положения центроидов. Некоторые такие положения могут приводить к довольно-таки бредовым результатам. Поэтому k-средних лучше проводить несколько раз подряд. Кстати, если вы при этом каждый раз получаете разные результаты, стоит задуматься о смене количества групп.

НЕМАЛОВАЖНО ЗНАТЬ!
Метрики расстояний

Конкретные результаты кластерного анализа во многом зависят от того, какую *метрику расстояния* вы выбрали. А их существует несколько. Самая простая из них — *эвклидово* — есть просто кратчайший путь между двумя точками.



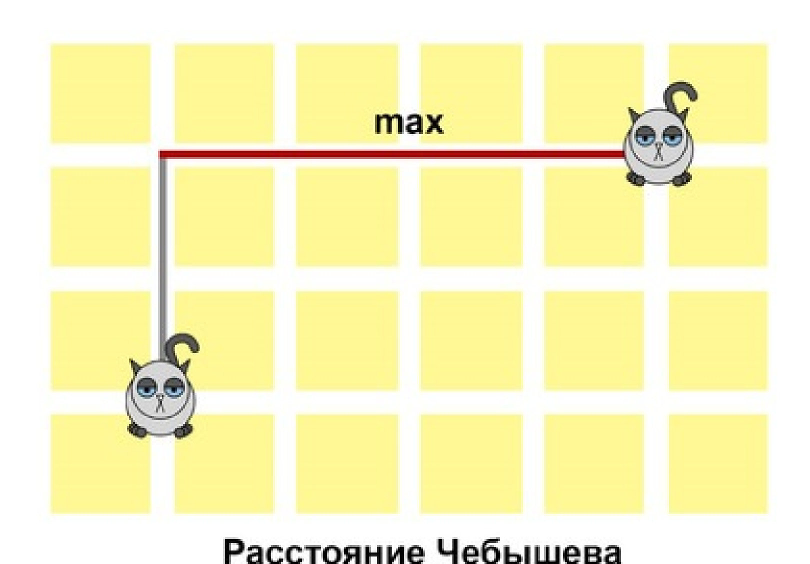
Иногда вместо него используют так называемое *Манхэттенское расстояние*. Названо оно было в честь Манхэттена, а точнее — в честь его планировки. Прогуливаясь по Манхэттену, вы не можете попасть из точки А в точку Б по кратчайшему пути. Если только вы не можете проходить сквозь стены, вам обязательно придется идти вдоль его параллельно-перпендикулярных улиц.



Манхэттенское расстояние

Заметим, что синий и красный пути абсолютно одинаковы. Манхэттенское расстояние лучше использовать в случаях, если вы подозреваете, что в вашей выборке есть выбросы.

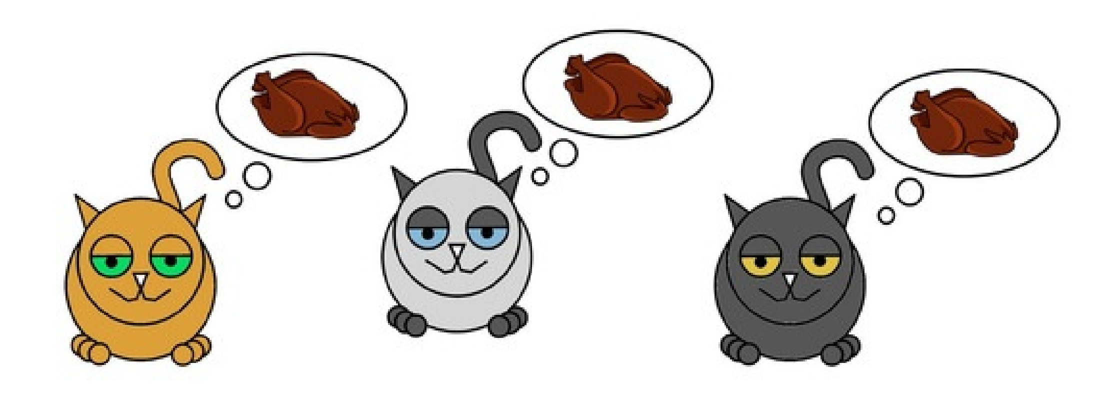
Последняя наиболее часто используемая метрика — это *расстояние Чебышева*. Она немного похожа на Манхэттенское расстояние. Но только чуть-чуть. Потому что его можно определить как максимальное расстояние, которое котику нужно будет пройти вдоль одной улицы.



Расстояние Чебышева

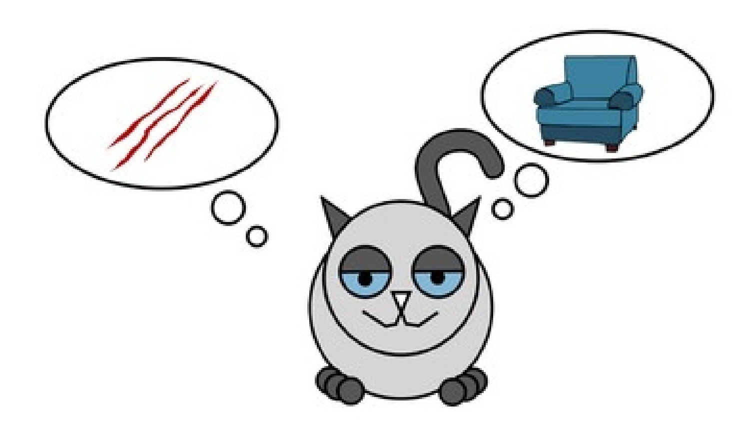
Глава 14.
О котиковом характере
или основы факторного анализа

Безусловно, каждый котик — уникальная и сложная личность. У него есть индивидуальные желания и предпочтения, а также собственный взгляд на мир и свое место в нем. Впрочем, некоторые психологические особенности (например, любовь к еде) являются общими для всех котиков.

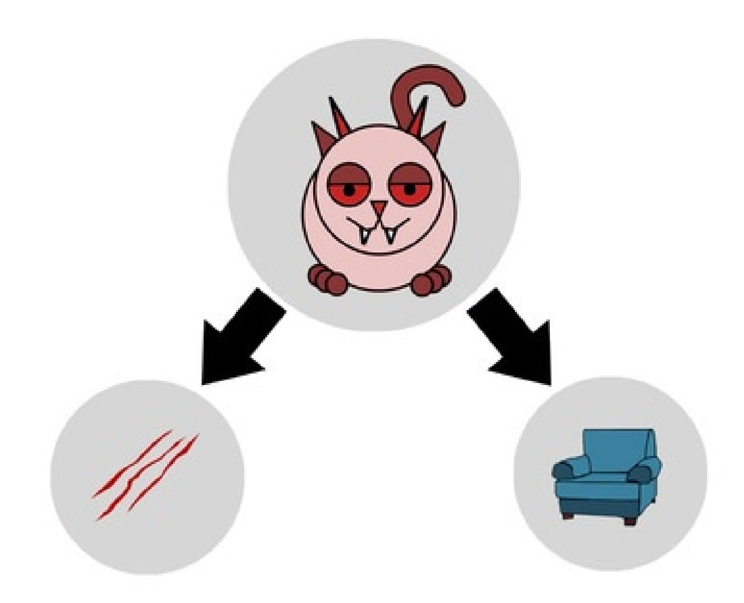


Однако, к большому сожалению, в отличие от всяких внешних признаков (к примеру таких, как размер или пушистость), психологические особенности не так просто измерить, поскольку их нельзя увидеть. И потому мы нуждаемся в специальных методах для их выявления.

В качестве примера вспомним, что большинство котиков склонны точить когти о диван и время от времени царапать своих хозяев. При этом мы наблюдаем линейную положительную взаимосвязь между этими явлениями — котики, которые дерут большее количество диванов, склонны оставлять большее количество царапин.



Глядя на эту взаимосвязь, мы можем предположить, что за этими склонностями стоит некоторая скрытая причина, которая вполне может являться особой чертой котикового характера. Назовем ее царапучестью. Чем выше царапучесть, тем больше котики склонны царапать диваны и людей.

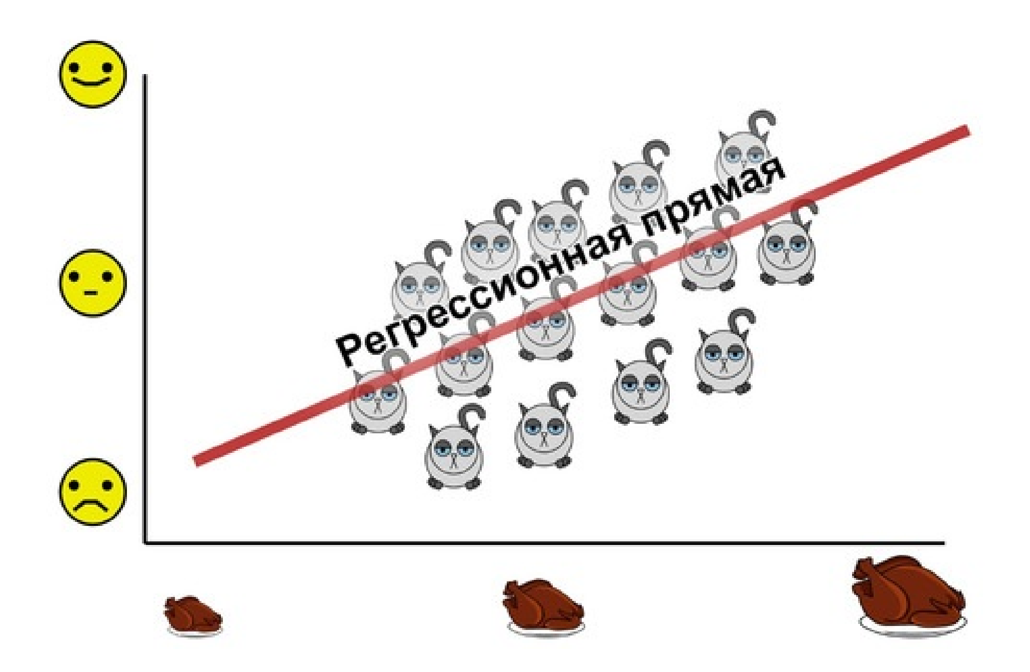


Выявить такие скрытые причины (или *факторы*) помогает *факторный анализ*, который проходит в несколько этапов. Во-первых, рассчитывается корреляционная матрица между всеми переменными, которые вы замерили: размером, количеством еды, склонностью царапать людей и т. д.

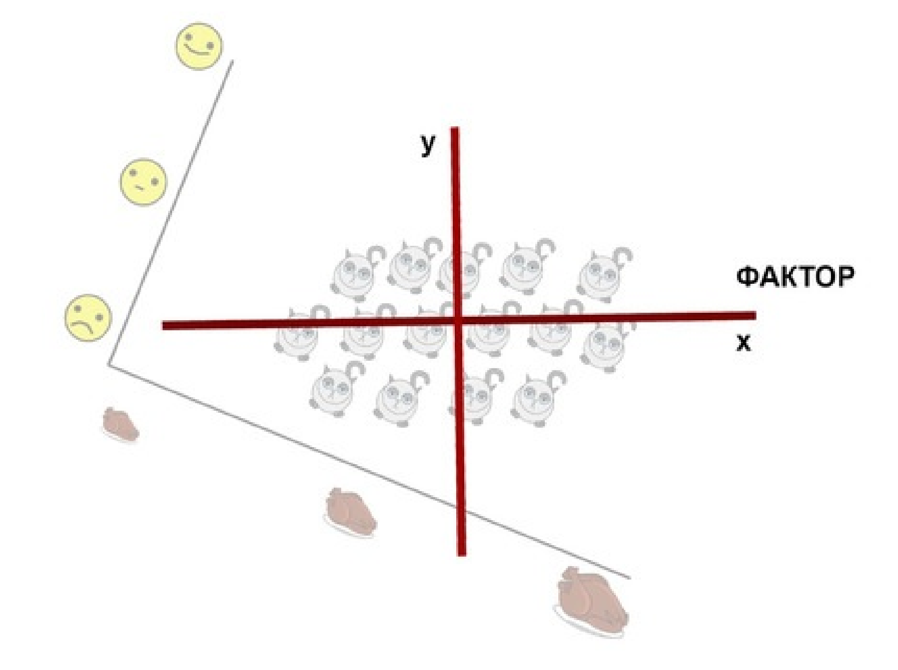
	1	0,9	0,2	0,3
	0,9	1	0,1	0,2
	0,2	0,1	1	0,8
	0,3	0,2	0,8	1

Во-вторых, переменные, которые коррелируют между собой, заменяются факторами. Чтобы понять, как это происходит, обратимся к рисунку.

На нем уже знакомая нам линейная взаимосвязь, которая описывается регрессионной прямой. Давайте теперь повернем наш рисунок таким образом, чтобы эта прямая лежала по горизонтали, и проведем прямую, перпендикулярную регрессионной.



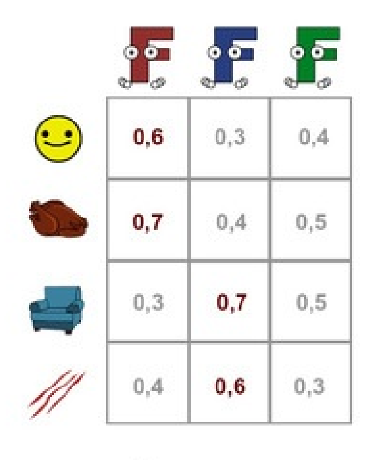
У нас получилась новая система координат. При этом большая часть котиков лежит вдоль оси X. Эта ось и будет являться фактором, заменяющим как количество поглощаемой пищи, так и котиковое счастье.



В итоге мы получаем вот такую таблицу, которая называется *факторной матрицей*. В каждой ячейке такой таблицы — коэффициент корреляции между одним из факторов и конкретной переменной. Называется он факторной нагрузкой. Сумма коэффициентов корреляции для каждого фактора называется *собственным значением*.

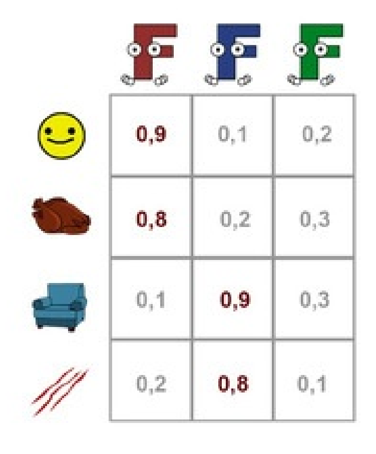
	F	F	F
	0,6	0,3	0,4
	0,7	0,4	0,5
	0,3	0,7	0,5
	0,4	0,6	0,3
Собственное значение Σ	2,0	2,0	1,8

Далее происходит так называемая процедура *вращения*. Цель ее заключается в том, чтобы большие коэффициенты корреляции в факторной матрице стали еще больше, а маленькие — еще меньше. Это значит, что каждый фактор будет связан только с определенной группой переменных и ни с какими другими.



	0,6	0,3	0,4
	0,7	0,4	0,5
	0,3	0,7	0,5
	0,4	0,6	0,3

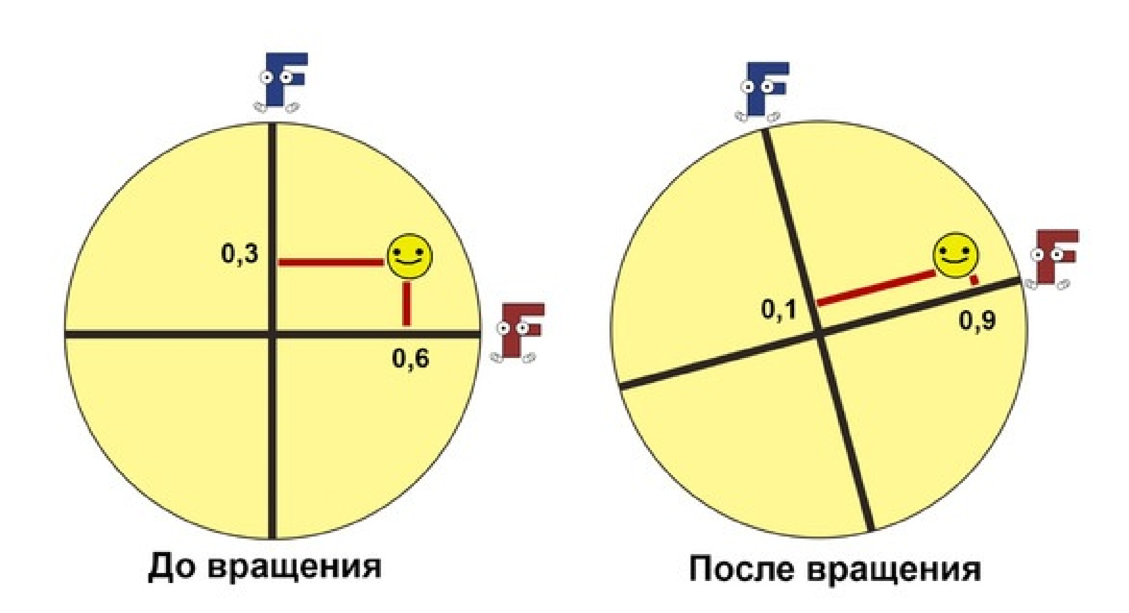
До вращения



	0,9	0,1	0,2
	0,8	0,2	0,3
	0,1	0,9	0,3
	0,2	0,8	0,1

После вращения

Чтобы прояснить, как работает вращение, также обратимся к рисунку. На нем изображена переменная «Счастье», которая коррелирует с первым и вторым факторами. Координаты «Счастье» — это коэффициенты корреляции между ним и факторами.

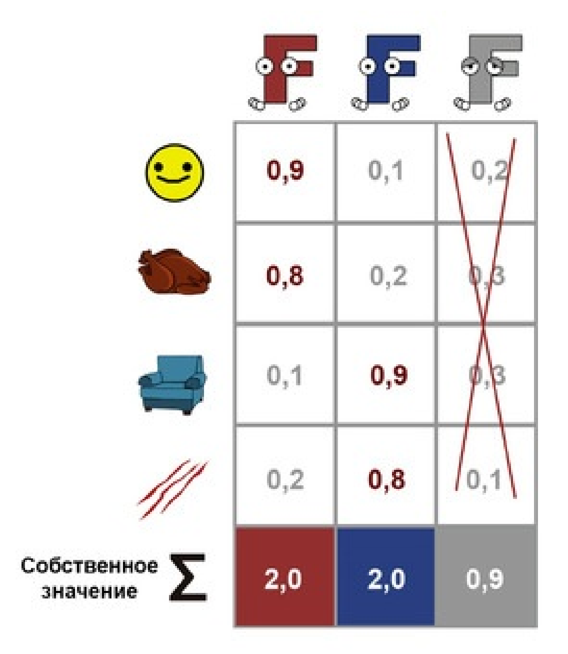


До вращения	После вращения

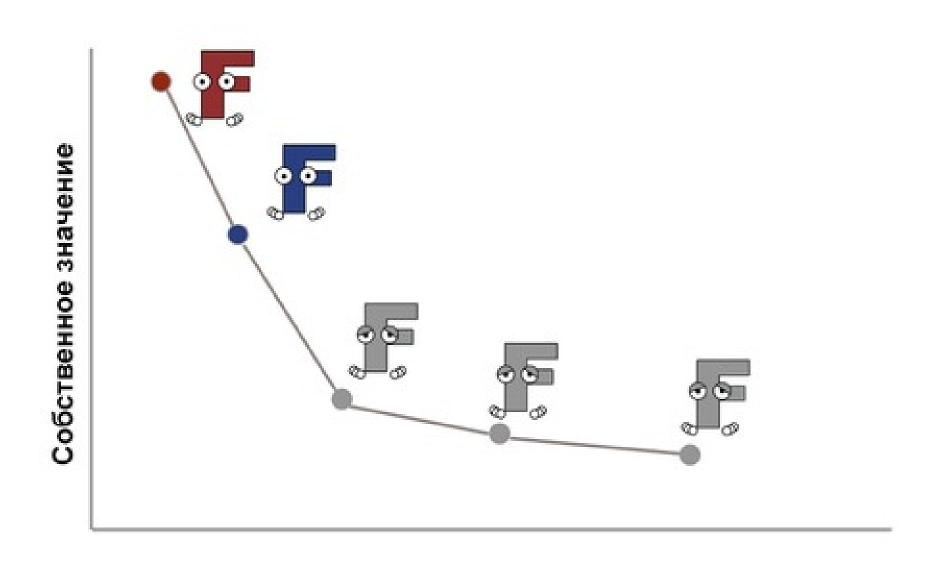
Если мы будем вращать окружность против часовой стрелки, то координаты «Счастья» будут меняться. Соответственно, оно будет больше коррелировать с первым фактором и меньше — со вторым.

Вращение бывает двух видов — *ортогональное* и *косоугольное*. В первом случае получившимся факторам запрещается коррелировать между собой, а во втором — нет.

Предпоследняя процедура — это отсеивание лишних факторов, которые слабо связаны с первоначальными переменными. Для этого существует два способа. Первый (называемый *критерием Кайзера*) заключается в том, что мы отбраковываем все факторы с собственным значением ниже 1.



Второй способ называется *методом каменистой осыпи* (или *критерием Кеттелла*). Для того чтобы им воспользоваться, необходимо построить график собственных значений. На горизонтальной оси этого графика располагаются факторы, а на вертикальной — их собственные значения. На определенной точке этого графика происходит перегиб. И все факторы, которые находятся за этой точкой, отсеиваются.



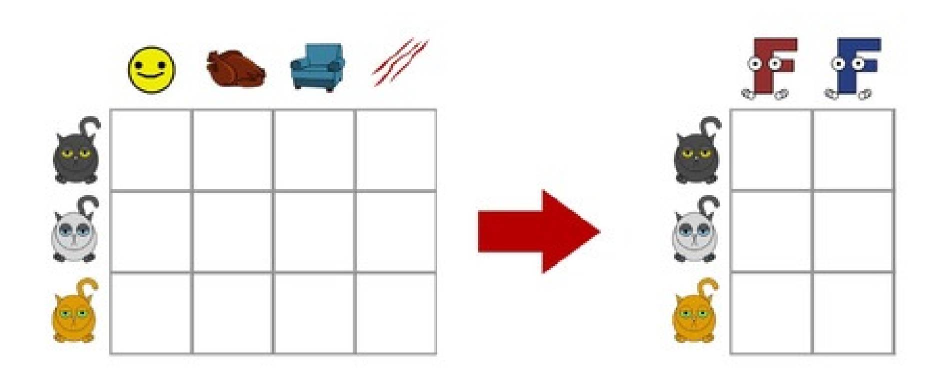
И наконец последний шаг — это придумать название получившимся факторам. Этот шаг является довольно нетривиальным — подчас он вызывает наибольшие затруднения. Но если вы успешно преодолеете его, то у вас на руках может оказаться довольно неплохая структурная модель котикового характера. В нашем случае первый фактор будет называться «жизнерадостностью», а второй — «царапучестью».

НЕМАЛОВАЖНО ЗНАТЬ!
Применение факторного анализа

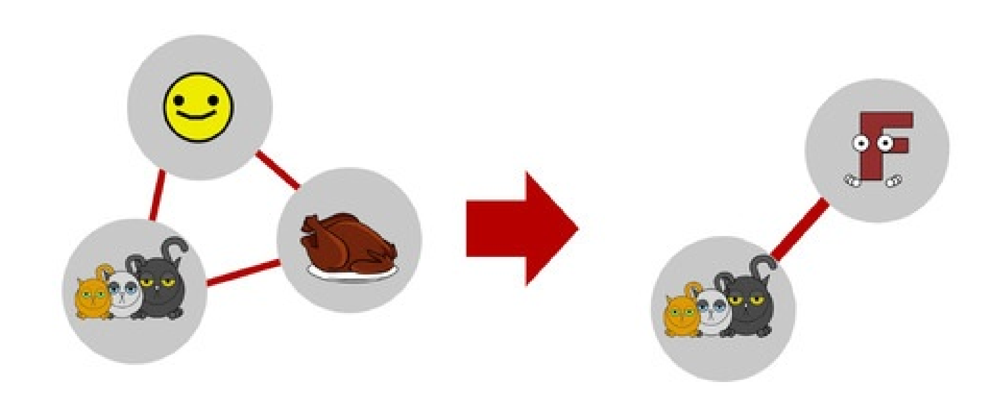
Изначально факторный анализ был разработан психологами для изучения способностей

и личностных качеств. Однако со временем область применения данного метода существенно расширилась.

Первая большая проблема, которую позволяет решить факторный анализ, это сокращение количества переменных. Как правило, серьезные исследования подразумевают сбор большого количества данных. Настолько большого, что в них бывает очень трудно разобраться. В этом случае факторный анализ позволяет уменьшить их количество за счет замены изначальных переменных факторами.



Вторая задача, требующая применения факторного анализа, это устранение мультиколлинеарности из регрессионных моделей. Напомним, что эта проблема заключается в том, что если две или более переменные взаимосвязаны между собой, результаты регрессионного анализа будут крайне ненадежными. Поэтому такие переменные требуется удалить из анализа. И один из путей — это замена таких переменных факторами.



Заключение

Ну вот и все. Ну, может, конечно, и не все: статистика все-таки гораздо богаче, и многое осталось за бортом. Но пока все. Потому что если объяснять совсем все, то пропадает интерес. А интерес — движущая сила в познании любого предмета. Да и потом, совсем все не объяснишь.

А так, мы рассмотрели самые базовые методы, которыми пользуются статистики для анализа данных. Мы прошлись по описательной статистике, рассмотрели меры различий и меры связи, познакомились с регрессионным и дискриминантным анализами, а также разобрались, как работают методы кластеризации и для чего используется факторный анализ. В общем, немало.

Надеюсь, что статистика стала вам ближе. Надеюсь, что страх и недоверие, если они и были, то прошли. Надеюсь, что вы заметили ту внутреннюю красоту, которая присуща этой дисциплине.

А в общем, надеюсь, что вам понравилось.

С уважением
Савельев Владимир

Приложение 1. Коротко о главном

В данном разделе кратко представлены методы, рассмотренные в книге, а также примеры их применения на практике. На этот раз без картинок и почти без котиков.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ, НЕОБХОДИМЫЕ ДЛЯ ПОНИМАНИЯ МАТЕРИАЛА

Генеральная совокупность — группа объектов, которые вам интересны как исследователю. В книге — все котики как биологический вид.

Выборка — часть генеральной совокупности, доступная для исследования. Статистики стремятся к тому, чтобы результаты, полученные на выборках, были верны и для генеральной совокупности. В книге описывается как котики, которых мы непосредственно измеряем.

Связанные выборки — ситуация, при которой любому объекту из первой выборки соответствует ровно один объект из второй. Можно сказать, что они образуют неразрывную пару (а в более сложных случаях — тройку, четверку и т. д.). В книге — котики до и котики после приема лекарства.

Наблюдение — измеренный объект. Котик.

Переменные — свойства объектов, которые поддаются измерению. В книге — котиковое счастье, здоровье, размер и т. д.

Значение переменной — степень выраженности того или иного свойства у конкретного объекта. Иными словами — насколько данный котик здоров, сыт и счастлив.

МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Используются, когда вам нужно отразить наиболее типичные значения, присутствующие в вашей выборке.

Состав:

1. Мода — наиболее часто встречающееся значение.

2. Медиана — середина упорядоченного ряда значений.

3. Среднее арифметическое — сумма значений, деленная на их количество.

Пример: *определение наиболее типичной зарплаты в нашей стране можно осуществлять по двум показателям — среднему арифметическому и медиане. Первая определяется как количество денег, деленное на количество людей, а второе — как зарплата человека, стоящего ровно посередине между самым бедным и самым богатым. Как правило, эти значения различаются — средняя зарплата выше медианной. И чем это различие больше, тем выше социальное неравенство в обществе.*

МЕРЫ ИЗМЕНЧИВОСТИ

Используются, когда нужно отразить степень разброса значений относительно меры центральной тенденции.

Состав:

1. Размах — разность между максимальным и минимальным значениями.

2. Дисперсия — сумма квадратов отклонений, деленная на их количество. *Отклонение* — это разность между средним арифметическим и конкретным значением. Дисперсии для генеральной совокупности и для выборки вычисляются по разным формулам.

3. Стандартное отклонение — корень из дисперсии.

Пример: *предположим, вы владеете заводом, который выпускает гвозди. Для любого*

массового производства необходимо, чтобы изделия полностью соответствовали некоторому стандарту. Например — длина ваших гвоздей должна быть ровно 10 см. Однако на практике всегда существуют некоторые отклонения от этого стандарта (например 10,2 или 9,7 см). Меры изменчивости позволяют оценить величину этих отклонений. Если стандартное отклонение длины превышает некоторое критическое значение, то ваша продукция не соответствует стандарту, а следовательно — не является качественной.

МЕРЫ РАЗЛИЧИЙ ДЛЯ НЕСВЯЗАННЫХ ВЫБОРОК

Позволяют определить различия между двумя несвязанными выборками. Наличие значимых различий по определенному признаку позволяет с некоторой уверенностью говорить о том, что генеральные совокупности также различаются. Эти методы делятся на параметрические и непараметрические. Первые желательно использовать только тогда, когда ваши данные удовлетворяют следующим требованиям.

1. Данные представлены в метрической шкале. Иными словами, признаки должны быть представлены в определенных единицах измерения (см, кг, сек. и т. д.)
2. Большое число наблюдений (от 30, но лучше более 100).
3. Распределение значений признаков приблизительно соответствует нормальному.
4. Отсутствуют выбросы (значения, на порядок отличающиеся от среднего).

Непараметрические меры различий работают и без этих допущений. Наиболее часто используемые меры различий представлены в таблице.

Вид	Две выборки	Три и более выборок
Параметрические	t-критерий Стьюдента для несвязанных выборок	Дисперсионный анализ
Непараметрические	U-Манна Уитни	H-Краскелла-Уоллеса

Пример: *предположим, что вы выращиваете помидоры, и вам необходимо определить, какой из двух сортов демонстрирует лучшую урожайность. Чтобы это сделать, вам необходимо подсчитать количество помидоров при каждом кусте и занести эту информацию в таблицу. Дальше вы применяете к этим данным t-критерий Стьюдента и по нему судите о наличии различий между сортами. Если сортов больше двух, то ваш выбор — дисперсионный анализ с последующим сравнением с помощью специальных post-hoc-критериев.*

МЕРЫ РАЗЛИЧИЙ ДЛЯ СВЯЗАННЫХ ВЫБОРОК

Позволяют определить различия между двумя связанными выборками. Также делятся на параметрические и непараметрические:

Вид	Две выборки	Три и более выборок
Параметрические	t-критерий Стьюдента для связанных выборок	Дисперсионный анализ для повторных измерений
Непараметрические	T-Вилкоксона	Критерий Фридмана

Пример: *Представим, что вы преподаватель курсов повышения квалификации, и вам интересно узнать, вынесли ли ваши слушатели что-нибудь полезное с занятий. Чтобы это сделать, вам необходимо разработать некоторый проверочный тест и раздать его слушателям до начала занятий и после их окончания. Т-критерий Вилкоксона позволит вам проверить, стали ли слушатели лучше знать ваш предмет. Если же вы провели несколько таких измерений, то ваш вариант — это критерий Фридмана.*

МЕРЫ СВЯЗИ

Данный класс критериев (называемых также коэффициентами корреляции) позволяет найти взаимосвязь между переменными. Математически взаимосвязь — это совместное изменение переменных.

Если она положительна и равна 1, то увеличение значения первой переменной сопровождается увеличением значения второй. Если она отрицательна (-1), то высокое значение первой переменной сопровождается низким значением второй. Коэффициент корреляции, равный 0, обозначает отсутствие взаимосвязи.

Самыми популярными коэффициентами корреляции являются r Пирсона (параметрический) и р Спирмена (непараметрический).

Пример: *вы решили провести психологическое исследование и выяснить, существует ли взаимосвязь между интеллектом и уровнем дохода. Для этого вам необходимо найти группу испытуемых, измерить их интеллект, узнать их среднемесячный доход и найти коэффициент корреляции. Если он высок и положителен, то более интеллектуальные люди получают больше денег.*

Если вы получили подобный результат, необходимо быть очень внимательными при его интерпретации. Поскольку равновероятными могут быть следующие варианты.

Более умные люди получают работу с более высоким заработком.

Высокий доход позволяет больше времени уделять саморазвитию в целом и развитию интеллекта в частности.

Существует неизвестная переменная (фактор), обусловливающая эту взаимосвязь.

Взаимосвязь является случайным совпадением.

РЕГРЕССИОННЫЙ АНАЛИЗ

Данная группа методов позволяет построить функциональную математическую модель — уравнение, которое помогает предсказать значение некоторой целевой переменной, используя значения ряда переменных, называемых предикторами.

Наиболее распространенными методами регрессионного анализа являются линейная и логистическая регрессии. Линейная регрессия позволяет предсказать точное количественное значение некоторой переменной, представленной в метрической шкале. Логистическая регрессия позволяет предсказать вероятность принадлежности объекта к тому или иному классу.

Пример: *предположим, вы управляете сетью розничных магазинов и хотите получить представление о том, какие факторы влияют на ежемесячную выручку в этих магазинах. Для этого вы должны замерить все возможные факторы, которые, по вашему мнению, могут на эту выручку повлиять: количество людей, посещающих магазин, число сотрудников на кассах, наличие на полках определенного товара и т. д. Затем необходимо построить линейную регрессию, указав в качестве целевой переменной выручку с этих магазинов, а в качестве предикторов — все, что вы замерили.*

Получив регрессионную модель, вы сможете не только посмотреть, какие факторы влияют на продажи, но и предсказать, какую выручку будет получать магазин при

определенных условиях.

Если вы немного скорректируете вашу задачу и примените метод логистической регрессии, то вы сможете узнать условия, при которых ваш магазин будет прибыльным или убыточным.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Дискриминантный анализ во многом похож на логистическую регрессию. Задачу, которую он решает, можно приблизительно сформулировать так: по каким переменным я могу отнести конкретный объект в тот или иной класс.

Пример: *предположим, вы проводите медицинское исследование и хотите узнать, по каким диагностическим показателям можно отличить больного человека от здорового. Для этого вы берете группы заведомо здоровых и больных людей и замеряете у них всех возможных «подозреваемых». После этого необходимо провести дискриминантный анализ, который и выявит систему показателей, по которым можно установить конкретный диагноз.*

КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ позволяет разбить ваши объекты на классы. При этом число классов может быть заранее неизвестным, либо вы точно знаете их количество. В первом случае ваш выбор — это метод иерархической кластеризации, который последовательно объединяет объекты в группы, основываясь на расстоянии между ними. Для второго случая необходим метод k-средних, который группирует ваши объекты вокруг так называемых центроидов.

Пример: *представим себе, что вы занимаетесь онлайн-продажами, и вам необходимо выделить категории клиентов, для того чтобы организовать более эффективную таргетированную рекламу. Чтобы это сделать, вы можете запустить на своем сайте небольшой опросник и, собрав некоторые данные о посещаемости тех или иных страниц, провести кластерный анализ. Если у вас есть некоторые предположения о том, какие именно категории клиентов заходят к вам на сайт, ваш выбор k-средних. Если таких предположений нет — то можно обойтись иерархической кластеризацией.*

ФАКТОРНЫЙ АНАЛИЗ

Факторный анализ позволяет сократить количество переменных, заменив их набором факторов. Кроме того, он может являться предварительной процедурой перед проведением регрессионного анализа в случае, если ряд предикторов коррелирует между собой.

Пример: *предположим, вы разрабатываете батарею психологических тестов, предназначенную для диагностики способностей у школьников. После того, как вы составили ряд задач, а также провели их на выборке учащихся, вам необходимо будет провести факторный анализ. Если высокий балл по одной задаче, как правило, сопровождается высоким баллом по другой задаче, значит, за ними скорее всего стоит некоторый общий фактор. Этот фактор и будет указывать на уровень развития той или иной способности.*

Приложение 2.
Работа в статистических пакетах

На сегодняшний день существует огромное количество программных продуктов, которые позволяют работать если не со всеми, то во всяком случае с большинством методов, о которых рассказывается в книге. В первом приближении их можно поделить на два класса: те, в которых все команды задаются с помощью текстового ввода (например *R* и *Python*), и те, где конкретный метод выбирается с помощью меню. Поскольку рядовой пользователь достаточно редко имеет дело с командной строкой, мы остановимся только на втором классе программ. Самыми популярными из них можно считать следующие.

1. IBM SPSS — мощный пакет, способный справиться с абсолютным большинством статистических задач. Является платным, однако существует и бесплатная 14-дневная версия.

2. StatSoft Statistica — главный конкурент *SPSS* на отечественном рынке. Также является коммерческим продуктом.

3. R-commander — графический интерфейс для языка программирования R. Как и сам R, распространяется бесплатно.

4. PSPP — бесплатный аналог *SPSS* со схожим интерфейсом.

5. Microsoft Excel с надстройкой «Анализ данных». Как ни странно, позволяет делать довольно много интересных вещей. Но его интерфейс не является типичным для статистических программ.

Здесь мы рассмотрим, как работать с *SPSS*. Однако многие вещи, о которых пойдет речь ниже, подходят и для других статистических пакетов. В частности, для любой статистической программы с меню характерна вот такая последовательность работы:

1. Вбить данные в таблицу;
2. Найти нужный метод;
3. Выбрать переменные для анализа;
4. Отметить необходимые опции;
5. Нажать «ОК»;
6. Проинтерпретировать результаты.

При этом первый, пятый и шестой шаги практически полностью идентичны. В частности, когда вы вбиваете данные в таблицу, абсолютное большинство пакетов следуют следующему правилу:

«По строкам — объекты, по столбцам — переменные».

При этом если у вас присутствуют несвязанные выборки, то этот факт кодируется отдельной переменной, которая обозначает принадлежность объекта к той или иной группе (например, 0 — котик и 1 — кошечка). В свою очередь каждая связанная выборка обозначается отдельной переменной (например, «Размер до» и «Размер после»).

Объект	Пол (0-котик, 1-кошечка)	Размер до (см)	Размер после (см)
Барсик	0	62	64
Мурзик	0	67	68
Тишка	0	65	67
Дуся	1	57	60
Муся	1	54	55
Мурка	1	52	54

Остальные шаги отличаются некоторыми нюансами, которые зависят как от пакета, так и от метода. В частности, в *SPSS* выбор переменных осуществляется с помощью переноса их в отдельные поля, а, допустим, в *Statistica* — простым выделением мыши.

Итак, ниже будут приведены алгоритмы работы в программе *IBM SPSS Statistics 24* (пробная русская версия с официального сайта). Они будут состоять из четырех разделов:

1. КАК НАЙТИ, в котором указывается путь к конкретному методу. Он всегда начинается с верхнего меню (там, где «Файл», «Изменить» и т. д.);

2. ЧТО ВВОДИТЬ — что необходимо сделать для проведения анализа.

3. ДОПОЛНИТЕЛЬНЫЕ ОПЦИИ, которые позволяют приспособить метод под вашу конкретную задачу.

4. КУДА СМОТРЕТЬ — указание на таблицы и ячейки, в которых содержатся основные результаты анализа.

ОПИСАТЕЛЬНАЯ СТАТИСТИКА И ДИАГРАММЫ

Как найти: *Анализ —> Описательные статистики —> Частоты…*

Что вводить: Выделите переменные, которые вы хотите проанализировать, и с помощью стрелочки перенесите их в поле «*переменные*».

Дополнительные опции:

Статистики… — позволяет выбрать конкретные меры центральной тенденции и меры изменчивости.

Диаграммы… — позволяет выбрать диаграммы (круговую или столбчатую).

Формат… — позволяет отрегулировать, в каком виде будет выдаваться результат. Например, можно вывести результаты по каждой переменной по отдельности, а можно — вместе.

Куда смотреть: в таблицы с описательными статистиками и на диаграммы.

Т-КРИТЕРИЙ СТЬЮДЕНТА ДЛЯ НЕСВЯЗАННЫХ ВЫБОРОК

Как найти*:* *Анализ —> Сравнение средних —> Т-критерий для независимых выборок.*

Что вводить:

1. Переместите переменные, по которым хотите найти различия, в поле «*Проверяемые переменные*».

2. Переместите переменную, которая делит ваши объекты на группы (т. е. На несвязанные выборки), в поле «*Группировать по*».

3. Задайте группы, либо указав конкретные значения (например 0 и 1), либо обозначив некоторое пороговое, ниже которого будет одна группа, а выше — другая.

Дополнительные опции: ничего интересного.

Куда смотреть: смотрим в таблицу «*Критерий для независимых выборок*». Слева будет два важных столбца, обозначающих критерий равенства дисперсий Ливиня, который определяет, равны ли между собой дисперсии ваших выборок.

Если значимость больше 0,05, то они равны и вам дальше нужно будет смотреть в первую строчку («*Предполагаются равные дисперсии*»). Если меньше 0,05 — то во вторую («*Не предполагаются равные дисперсии*»).

Следующие столбцы — сам t-критерий Стьюдента. Если его значимость меньше 0,05 (столбец «*Знач. Двухсторонняя*»), то средние значения ваших выборок различаются. Если же больше 0,05, то таких различий обнаружено не было.

Если вы хотите узнать, у какой группы соответствующий показатель больше, смотрите в таблицу «С*татистика группы*» (столбец «*Средние*»).

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Как найти: *Анализ —> Общая линейная модель —> ОЛМ-одномерная.*

Что вводить:

1. Переместите переменную, по которой хотите найти различия, в поле «*Зависимая переменная*».

2. Переместите переменные, которые делят ваши объекты на группы (т. е. на несвязанные выборки), в поле «*Фиксированные факторы*».

Дополнительные опции:

Апостериорные — позволяет вычислить различные post-hoc-критерии.

Параметры — разные дополнительные критерии. Как правило, нас интересуют описательные статистики. Также весьма полезным может быть график средних.

Куда смотреть: нас интересуют два последних столбца таблицы «*Критерии межгрупповых эффектов*» — «*F*» и «*Значимость*». Эти параметры есть при каждом факторе. Если «*Значимость*» меньше 0,05 — фактор влияет на переменную.

Если вы включили post-hoc-критерии, то найти их можно в таблице «*Множественные сравнения*». Средние показатели по каждой группе вы сможете найти в таблице «*Описательные статистики*».

МНОГОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Как найти: *Анализ —> Сравнение средних —> Однофакторный дисперсионный анализ.*

Что вводить:

1. Переместите переменные, по которым хотите найти различия, в поле «*Список зависимых переменных*».

2. Переместите переменную, которая делит ваши объекты на группы (т. е. на несвязанные выборки), в поле «*Фактор*».

Дополнительные опции:

Апостериорные — позволяет вычислить различные post-hoc-критерии.

Параметры — разные дополнительные критерии. Как правило, нас интересуют описательные статистики. Также весьма полезным может быть график средних.

Куда смотреть: смотрим на два последних столбца таблицы «*ANOVA*» — «*F*» и «*Значимость*». Если «*Значимость*» меньше 0,05 — фактор влияет на переменную.

Если вы включили *post-hoc*-критерии, то найти их можно в таблице «*Множественные сравнения*». Средние показатели по каждой группе вы сможете найти в таблице «*Описательные статистики*».

U-КРИТЕРИЙ МАННА-УИТНИ

Как найти: *Анализ —> Непараметрические критерии —> Устаревшие диалоговые окна -> Для двух независимых выборок.*

Что вводить:

1. Переместите переменные, по которым хотите найти различия, в поле «*Список проверяемых переменных*».

2. Переместите переменную, которая делит ваши объекты на группы (т. е. на несвязанные выборки), в поле «*Группировать по*».

3. Задайте группы, указав конкретные значения (например 0 и 1).

Дополнительные опции: если хотите, можете посмотреть различия по другим критериям.

Куда смотреть: смотрим в таблицу «*Статистические критерии*». Сам критерий U Манна-Уитни находится в одноименной строчке. P-уровень значимости можно найти в строчке «*Асимптотическая значимость (2-сторонняя)*». Если он меньше 0,05, ваши выборки значимо различаются. Если же больше 0,05, то таких различий обнаружено не было.

H-КРИТЕРИЙ КРАСКЕЛЛА-УОЛЛЕСА

Как найти: *Анализ —> Непараметрические критерии —> Устаревшие диалоговые окна -> Для K независимых выборок.*

Что вводить:

1. Переместите переменные, по которым хотите найти различия, в поле «*Список проверяемых переменных*».

2. Переместите переменную, которая делит ваши объекты на группы (т. е. на несвязанные выборки), в поле «*Группировать по*».

3. Задайте группы, указав диапазон их значений. Например от 1 до 3 в случае, если у вас 3 группы.

Дополнительные опции: ничего интересного.

Куда смотреть: смотрим в таблицу «*Статистические критерии*». Абсолютное значение критерия скрывается в строчке «*Хи-квадрат*». Если «*Асимптотическая значимость меньше 0,05*», то влияние фактора можно считать значимым.

T-КРИТЕРИЙ СТЬЮДЕНТА ДЛЯ СВЯЗАННЫХ ВЫБОРОК

Как найти: *Анализ —> Сравнение средних —> T-критерий для парных выборок.*

Что вводить: переместите пары переменных, обозначающих связанные выборки в поле «*Парные переменные*».

Дополнительные опции: ничего интересного.

Куда смотреть: смотрим в таблицу «*Критерий парных выборок*» на последние столбцы. «*T*» — значения критерия, а «*Знач. (двухсторонняя)*» показывает p-уровень значимости. Если он меньше 0,05 — различия имеются.

Если вы хотите узнать, у какой группы соответствующий показатель больше, смотрите в таблицу «*Статистика парных выборок*» (столбец «*Среднее*»).

ДИСПЕРСИОННЫЙ АНАЛИЗ ДЛЯ ПОВТОРНЫХ ИЗМЕРЕНИЙ

Как найти: *Анализ —> Общая линейная модель —> ОЛМ-повторные измерения.*

Что вводить:

1. Задайте имя внутригруппового фактора, по которому разделяются ваши связанные выборки, число уровней (кол-во связанных выборок) и нажмите кнопку «*Добавить*».

2. Переместите переменные, обозначающие ваши связанные выборки, в поле

«*Внутригрупповые переменные*».

Дополнительные опции: если у вас имеются несвязанные выборки, то вы можете включить их в анализ, добавив соответствующую переменную в межгрупповые факторы.

В разделе «*Графики*» вы можете настроить выдачу графиков средних по каждому фактору.

Куда смотреть: смотрим в таблицу «*Критерии внутригрупповых эффектов*» (блок с названием внутригруппового фактора). Там — четыре критерия, у которых чаще всего одинаковые значения (столбец *F*). Если «*Значимость*» при них меньше 0,05, то связанные выборки различаются между собой.

Т-КРИТЕРИЙ ВИЛКОКСОНА

Как найти: *Анализ —> Непараметрические критерии —> Устаревшие диалоговые окна —> Для двух связанных выборок.*

Что вводить: переместите пары переменных, обозначающих связанные выборки, в поле «*Тестовые пары*».

Дополнительные опции: если хотите, можете посмотреть различия по другим критериям. Например, по критерию знаков.

Куда смотреть: смотрим в таблицу «*Статистические критерии*». Т-критерия Вилкоксона вы в ней не найдете — вместо него так называемая Z-статистика, рассчитанная на основе этого критерия. Ее вполне можно вставлять в вашу работу.

Р-уровень значимости можно найти в строчке *«Асимптотическая значимость (2-сторонняя)»*. Если он меньше 0,05, ваши выборки значимо различаются. Если же больше 0,05, то таких различий обнаружено не было.

КРИТЕРИЙ ФРИДМАНА

Как найти: *Анализ —> Непараметрические критерии —> Устаревшие диалоговые окна —> Для K связанных выборок.*

Что вводить: переместите переменные, обозначающие связанные выборки, в поле «*Проверяемые переменные*».

Дополнительные опции: ничего интересного.

Куда смотреть: смотрим в таблицу «*Статистические критерии*». Абсолютное значение критерия скрывается в строчке «*Хи-квадрат*». Если «*Асимптотическая значимость меньше 0,05*», то влияние фактора можно считать значимым.

КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ ПИРСОНА И СПИРМЕНА

Как найти: *Анализ —> Корреляции —> Парные.*

Что вводить:

1. Переместите переменные, между которыми вы хотите найти взаимосвязи, в поле «*Переменные*».

2. Выберите нужный коэффициент корреляции.

Дополнительные опции: ничего интересного.

Куда смотреть: программа выдаст вам корреляционную матрицу (таблица «*Корреляции*» или «*Непараметрические корреляции*»). Чтобы посмотреть в ней коэффициент корреляций между переменными А и Б, нужно найти строчку с переменной А и столбик с переменной Б и посмотреть, где они пересекаются.

Сверху будет коэффициент корреляции, а чуть ниже — уровень значимости (двухсторонний). Если он ниже 0,05, то связь между переменными действительно присутствует.

ЛИНЕЙНАЯ РЕГРЕССИЯ

Как найти: *Анализ —> Регрессия —> Линейная…*

Что вводить:

1. Переместите целевую переменную в поле «*Зависимая переменная*».
2. Переместите переменные-факторы в «*Независимые переменные*».

Дополнительные опции: на главном окне вы можете выбрать метод линейной регрессии. Как правило, «*Ввод*» и «*Пошагово*».

Нажав на кнопку «*Статистики*», вы сможете выбрать некоторые дополнительные коэффициенты, которые выдаст вам программа.

Куда смотреть: смотрим в таблицу «*Коэффициенты*». Там нас будут интересовать два столбца — «*B*» и «*Значимость*». В первом из них — регрессионные коэффициенты. Во втором — p-уровень значимости. Если он меньше 0,05, то данный фактор является значимым.

Вторая интересующая нас таблица — сводка для модели. Смотрим столбец «*Скорректированный R-квадрат*». В нем — коэффициент детерминации, который скажет, какой процент ваших данных объясняет модель. R-квадрат, равный 0,92, обозначает, что 92% ваших данных объясняется вашей моделью.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Как найти: *Анализ —> Регрессия —> Логистическая…*

Что вводить:

1. Переместите целевую переменную в поле «*Зависимая переменная*».
2. Переместите переменные-факторы в «*Ковариаты*».

Дополнительные опции: на главном окне вы можете выбрать метод логистической регрессии. По умолчанию установлен «*Ввод*» (или «*Enter*»).

Нажав на кнопку «*Параметры*», вы сможете выбрать некоторые дополнительные статистики и графики. Также я очень рекомендую поставить галочку в графе «*На последнем шаге*».

Куда смотреть: пролистываем вывод вниз (до Блок 1) и смотрим в таблицу «*Переменные в уравнении*». Интересуют нас два столбца: «*B*» и «*Значимость*». Первый содержит регрессионные коэффициенты. Второй — p-уровень значимости. Если он меньше 0,05, то данный фактор является значимым.

Вторая таблица — «*Сводка для модели*». Смотрим столбец «*R-квадрат Нэйджелкерка*». Этот коэффициент показывает, сколько процентов ваших данных объясняет полученная модель. R-квадрат, равный 0,92, обозначает, что 92% ваших данных объясняется вашей моделью.

И последнее — «*Таблица классификации*». Она позволяет сравнить, насколько результаты, предсказываемые моделью, совпадают с реальными.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Как найти: *Анализ —> Классификация —> Дискриминантный анализ.*

Что вводить:

1. Переместите переменную, делящую ваши объекты на группы, в поле «*Группировать по*». Далее — задайте диапазон, в котором находятся ваши группы (допустим от 1 до 3, если группы обозначаются как 1, 2 и 3).
2. Переместите остальные переменные в поле «*Независимые*».
3. Нажмите кнопку «*Статистики*» и отметьте «*Однофакторный дисперсионный анализ*».
4. Нажмите кнопку «*Классифицировать*» и отметьте «*Итоговая таблица*».

Дополнительные опции: на главном окне вы можете выбрать метод дискриминантного анализа («*Принудительное включение*» или «*Шаговый отбор*»).

В окне «*Статистики*» вы также можете выбрать «*Средние*», что даст описательную статистику по каждой из групп.

Куда смотреть: в таблице «Критерии равенства групповых средних» можно посмотреть, какие переменные значимо разделяют ваши объекты на группы (столбцы «*F*» и «*Значимость*»). Если значимость меньше 0,05, то разделяет.

Значения коэффициентов стандартизованной канонической дискриминантной функции можно найти в одноименной таблице (если это действительно необходимо).

Что касается меры качества, то таковой может служить таблица «*Результаты классификации*». В ячейках [0,0] и [1,1] находятся правильно классифицированные объекты, а в остальных — ошибочно определенные.

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Как найти: *Анализ —> Классификация —> Иерархическая кластеризация…*

Что вводить:

1.Переместите признаки, по которым ваши объекты будут распределяться на группы, в поле «*Переменные*».

2. В разделе «*Графики*» отметьте галочкой «*Дендрограмма*».

Дополнительные опции: нажав кнопку «*Статистики*», вы можете потребовать у компьютера вывести принадлежность объектов к кластерам на том или ином этапе кластеризации. Кроме того, у него можно затребовать матрицу расстояний между объектами (она же — «*Матрица близостей*»).

В разделе «*Метод*» вы можете выбрать способ выделения кластеров, а также меру расстояния.

Куда смотреть: на дендрограмме показана принадлежность объектов к тому или иному классу на всех этапах кластеризации.

Если же вы отметили соответствующую галочку, то вы можете посмотреть принадлежность объектов к кластеру на определенном этапе кластеризации в таблице «*Принадлежность к кластерам*».

К-СРЕДНИХ

Как найти: *Анализ —> Классификация —> Кластеризация К-средними.*

Что вводить:

1. Переместите признаки, по которым ваши объекты будут распределяться на группы, в поле «*Переменные*».

2. Выберите число кластеров.

3. В разделе «*Параметры*» отметьте «*Конечный кластер для каждого наблюдения*».

Дополнительные опции: ничего интересного.

Куда смотреть: из таблицы «*Принадлежность к кластерам*» можно увидеть, какой объект к какому кластеру принадлежит.

А в таблице «*Конечные центры кластеров*» расположены координаты каждого центроида.

ФАКТОРНЫЙ АНАЛИЗ

Как найти: *Анализ —> Снижение размерности —> Факторный анализ.*

Что вводить:

1. Переместите переменные, на основе которых будут выделяться факторы, в поле

«*Переменные*».

2. Нажмите на кнопку «*Вращение*» и выберите метод вращения (чаше всего «*варимакс*»).

Дополнительные опции: в разделе «*Извлечение*» можно выбрать метод извлечения, вывести график собственных значений или настроить количество факторов, которые выделятся по итогу.

Куда смотреть: результаты факторного анализа находятся в «*Повернутой матрице компонентов*». Там — коэффициенты корреляции между факторами и отдельными переменными.

Собственные значения факторов можно посмотреть в таблице «*Объясненная совокупная дисперсия*».

Приложение 3. Что еще посмотреть?

Если после прочтения данной книги вы заинтересовались статистикой, то было бы не лишним узнать, что еще можно посмотреть по данной тематике.

В первую очередь я бы рекомендовал курсы института биоинформатики на сайте www.stepik.org. А именно «Основы статистики» в трех частях, который ведут Анатолий Карпов, Иван Иванчей, Полина Дроздова и Арсений Москвичев. Там все просто, доходчиво и талантливо. А демонстрируемая глубина изложения встречается далеко не в каждом учебнике.

Второй источник, достойный упоминания — это «Статистика для всех» С. Бослаф. Единственное — она весьма недешёвая и её трудно найти. Содержание же выше самых похвал — подробно рассмотрены самые распространенные методы обработки данных, в том числе и специфические для медицины, экономики и бизнеса.

Также я достаточно часто захожу на портал знаний statistica.ru компании StatSoft. Местный электронный учебник хорош в качестве справочного пособия. Что касается самого анализа данных в системе Statsoft Statistica, то о нём можно узнать в учебнике Боровикова «Популярное введение в современный анализ данных в системе STATISTICA».

Если же вам приходится работать в SPSS — возьмите книгу А. Д. Наследова «IBM SPSS Statistics 20 и AMOS: Профессиональный статистический анализ данных». Там описано решение большинства типовых задач, с которыми приходится сталкиваться исследователю.

По статистическому языку R есть прекрасный курс на том же stepik.org. Ведут Анатолий Карпов и Иван Иванчей.

А вообще, самый главный источник знаний — это исследовательская работа. Решение практических задач способствует их усвоению и закреплению в гораздо большей степени, чем чтение книг. Поэтому если вы хотите освоить этот предмет — ищите достойные задачи, решение которых позволит сделать наш мир лучше и интереснее.

Благодарности

Здесь мне хотелось бы выразить благодарность людям, без которых издание книги было бы невозможным.

И в первую очередь спасибо тем, кто поверил в этот проект и вложился в него, став спонсорами на краудфандинговой площадке Boomstarter. Без них он так и остался бы просто красивой идеей. Ваша поддержка вдохновляла меня, а ответственность перед вами заставляла ежедневно работать над книгой, делая ее все лучше и лучше.

В особенности мне бы хотелось поблагодарить следующих спонсоров: Дмитрия Чумаченко, Елену Зеркаленкову, Анатолия Федоточкина, Леонида Тощева, Евгения Комоцкого, Ольгу Романову, Ивана Равового, Алексея Иванова (aviva24), Вадима Шмыгова и школу «Инфографика TUT», Максима Кравцова, Ирину Шафранскую, Сергея Черепанова, Владимира Волохонского, Александра Белоцерковского, Евгения Степанищева, Вячеслава Калошина и Игоря Мосягина. Их вклад был по-настоящему щедрым и позволил реализовать несколько интересных идей.

Среди них есть три человека, которых я знаю лично и которым я бы хотел выразить отдельную благодарность. В частности, благодаря Дмитрию Чумаченко в свое время я и занялся анализом данных. Именно его меткое замечание во время одного моего выступления на конференции подвигло меня на изучение этой дисциплины.

Взаимообмен идеями с Евгением Комоцким, моим коллегой и хорошим другом, помог мне сильно продвинуться в этой области. Спасибо ему за те удивительные и интересные задачи, которые нам вместе приходится решать.

Владимир Львович Волохонский был и остается для меня авторитетом в области сбора и обработки данных. Я горжусь тем, что он не только стал спонсором моего проекта, но и выступил в качестве эксперта для этой книги.

В связи с этим я бы хотел выразить огромную благодарность ему и другим экспертам, которые помогли сделать эту книгу гораздо лучше, чем она могла бы быть. Они нашли огромное количество ошибок и неточностей, опрометчиво допущенных мной, и не позволили мне ввести вас в заблуждение относительно некоторых важных тем.

Также спасибо Андрею Дмитриевичу Наследову, автору учебника «Математические методы психологического исследования», ставшего настольной книгой для многих психологов. Помимо экспертной оценки, которую он дал, я бы хотел поблагодарить его за отзыв о «Статистике и котиках». Этот отзыв придал мне уверенности в своих силах — я понял, что двигаюсь в правильном направлении.

Моя переписка с Анатолием Карповым достойна отдельной главы. Будучи психологом по образованию и преподавателем статистики в Институте биоинформатики, он, пожалуй, внес наибольший вклад в содержание книги. Огромное спасибо ему за консультации и экспертную оценку. И обратите внимание на курсы, которые он и его коллеги делают на сайте www.stepik.org. Они великолепны.

Помимо экспертов, значительный вклад в содержание книги внесли двое читателей блога: Алексей Русаков и Алексей Сотов. С последним, кстати, мы дружим уже много-много лет.

Спасибо администраторам групп «ВКонтакте», согласившихся опубликовать у себя новость о книге. Особая благодарность Исмаилу Алиеву за живой интерес и неоценимую помощь в продвижении проекта в социальных сетях.

Также я хотел бы поблагодарить людей, непосредственно работавших со мной над реализацией «Статистики и котиков»: Сысоеву Анну из компании Boomstarter, которая помогла

организовать краудфандинговую кампанию, и Марию Рявину из издательства Ridero за помощь в организации печати и доставки тиража до спонсоров. За обложку, кстати, спасибо Максиму Силенкову. А Александра Бахманова и Ирина Знаменская помогли скрыть мою орфографическую и пунктуационную безграмотность.

Особая благодарность — Корженевскому Юрию. Он стал настоящим ангелом-хранителем этого проекта. Он приложил руку буквально ко всему — начиная с оказания значительной финансовой поддержки, заканчивая поиском дизайнера для обложки. Но самое важное, что я от него получил, это правильные вопросы, заданные им в правильное время. Я многому научился, работая с ним.

Наконец, я бы хотел поблагодарить своих родных, друзей и коллег за моральную поддержку и безграничное терпение. Со мной реально было тяжело в эти месяцы.

И спасибо Виталине. Без нее я бы не справился.